Autobiographical Memory Impairments as a Transdiagnostic Feature of Mental Illness: A meta-analytic review of investigations into autobiographical memory specificity and overgenerality amongst people with psychiatric diagnoses

Tom J. Barry¹; David J. Hallford²; Keisuke Takano³;

¹Faculty of Social Sciences, University of Hong Kong, Pok Fu Lam, Hong Kong, Hong Kong
²School of Psychology, Deakin University, Melbourne, Australia
³Division of Clinical Psychology and Psychotherapy, Department of Psychology, Ludwig-Maximilians-University Munich, München, Germany

*Corresponding author: Tom J. Barry, A: Faculty of Social Sciences, Jockey Club Tower, Pok Fu Lam Road, Hong Kong. E: tjbarry@hku.hk; tom.j.barry@icloud.com; T: +852 3917 7457.

Author note. The data, their associated analysis scripts and other supplementary tables and figures are available online at https://osf.io/3rjuz/ or 10.17605/OSF.IO/3RJUZ.
Abstract

Decades of research has examined the difficulty that people with psychiatric diagnoses have in recalling specific autobiographical memories of events that lasted less than a day. Instead, they seem to retrieve general events that have occurred many times or which occurred over longer periods of time, termed overgeneral memory. We present the first transdiagnostic meta-analysis of memory specificity/overgenerality, and the first meta-regression of proposed causal mechanisms. A keyword search of Embase, PsycARTICLES and PsycINFO databases yielded 74 studies that compared people with and without psychiatric diagnoses on the retrieval of specific ($k = 85$) or general memories ($k = 56$). The majority of studies included participants with Major Depressive Disorder (~49%), Schizophrenia (~19%) and Posttraumatic Stress Disorder (~17%) with few studies involving other groups of participants (e.g., Anxiety Disorders (~5%)). Multi-level meta-analysis confirmed that people with psychiatric diagnoses typically recall fewer specific ($g = -0.864, 95\% \text{ CI}[-1.030, -0.698]$) and more general ($g = .712, 95\% \text{ CI}[0.524, 0.900]$) memories than diagnoses-free people. The size of these effects did not differ between diagnostic groups. There were no consistent moderators of effect size heterogeneity; effect sizes were not explained by methodological factors such as cue valence, or demographic variables such as participants’ age, or between-group differences in process variables (e.g., rumination). Deficits in autobiographical memory retrieval may be a transdiagnostic factor, but further research in underrepresented diagnostic groups, and with novel experimental manipulations of encoding and retrieval processes, is warranted before full transdiagnosticity and the processes underlying reduced specificity/overgenerality can be established.

Keywords: Episodic Memory; Depression; Trauma; Schizophrenia; Borderline Personality Disorder
Public Significance Statement

The data presented here show that people with a range of psychiatric diagnoses experience difficulty retrieving memories of specific autobiographical events from their past. Instead, people with diagnoses tend to retrieve memories of general events that have occurred on multiple occasions or which occur over an extended period of time. These findings indicate a specific cognitive target for intervention in psychiatric problems, especially given that existing treatments do not typically help people in this way. These findings also contribute to a growing body of literature that indicates similarities in cognitive deficits that cut across discrete psychiatric diagnoses.
Introduction

In 1986, Williams and Broadbent asked people who had recently attempted suicide to retrieve specific memories – those referring to discrete events that were shorter than a day in duration (e.g., *my daughter’s birthday party last week*) – in response to each of several positive and negative cue words. People who had attempted suicide retrieved fewer specific memories than a control group of non-suicide attempters (Williams & Broadbent, 1986). Williams and Dritschel (1992) subsequently showed that instead of retrieving specific autobiographical memories, suicide attempters tended to retrieve *general* memories, such as *categories* of events that occurred on multiple occasions (e.g., *my daughters’ birthdays*) or events which lasted for *extended* periods of time (e.g., *when my daughter was two*). Subsequent studies have reported similar patterns of reduced autobiographical memory specificity (rAMS) and overgeneral memory (OGM) in a range of different psychiatric populations. These studies were first summarised in the work of Williams et al. (2007). Subsequent quantitative and qualitative reviews have been conducted in specific populations with discrete psychiatric presentations (Beran et al., 2019; Berna et al., 2016; Liu et al., 2013; Ono et al., 2015). However, no study has yet meta-analysed the entire literature regarding autobiographical memory specificity/overgenerality across all studies that compare people with psychiatric diagnoses to diagnoses-free people.

Such an analysis is important given criticisms to the categorical or diagnostic approach to psychopathology (C. C. Conway et al., 2019; Kotov et al., 2017). In particular, there is tremendous comorbidity that exists among psychiatric diagnoses (McGrath et al., 2020) and a focus on the transdiagnostic processes that underlie a range of different disorders may advance our understanding of how psychopathology in general, emerges and is maintained over time (Van den Bergh et al., 2020). This is consistent with idiographic approaches such as the Research Domain Criteria (RDoC; Insel et al., 2010) and Hierarchical
Taxonomy of Psychopathology (Hi-TOP; C. C. Conway et al., 2019), through which rAMS and OGM would be conceptualised as a failure of cognitive processes that support good mental health. Analyses of possible transdiagnostic constructs are also important as we navigate through an era where traditional psychological treatments are no more effective than they were decades ago (Cristea et al., 2017; Johnsen & Friborg, 2015) and where there are calls for novel interventions that target transdiagnostic mechanisms known to be associated with disorder maintenance (Craske, 2018).

Autobiographical memory specificity/overgenerality represents one such transdiagnostic mechanism, especially given early evidence of its presence across a range of diagnoses (Williams et al., 2007), meta-analytical evidence that it can lead to a maintenance or worsening of symptoms over time (Hallford et al., 2021) and evidence that interventions that target it can also influence symptoms, albeit temporarily (Barry et al., 2019). It is also possible that problems with autobiographical memory specificity may interfere with other seemingly unrelated therapies, such as cognitive-behavioural therapy, in so far as these therapies are themselves autobiographical experiences that must be retrieved if treatment gains are to be realised (Harvey et al., 2014; Hitchcock, Rudokaite, et al., 2019). As such, it is imperative that we provide an up-to-date quantitative analysis of the presence of rAMS/OGM across the various clinical presentations.

Given previous evidence (Williams et al., 2007), we might expect that pooled effect sizes for differences between psychiatric participants and their diagnoses-free counterparts in their retrieval of specific or general memories would be significantly different from zero. We might also expect that the size of these effects would vary between studies and that this variability might be explained by moderating factors related to study methodology (e.g., the duration of time given to participants to retrieve their memories), the population under examination (e.g., age) or differences between participants in other variables that are
theorised to be involved in memory retrieval. A broad, updated, analysis of the entire literature and the inclusion of many more studies than if only one disorder group were included, therefore allows us to explore the transdiagnostic nature of both rAMS and OGM and should be adequately powered to explore the moderating factors that explain why some groups of people experience particularly significant problems with their autobiographical memory.

One such group of moderators regards methodology and the measurement of specificity/generality. The majority of studies in this area use the Autobiographical Memory Test (AMT; Williams & Broadbent, 1986). The AMT typically gives participants 60 seconds to recall autobiographical memories in response to each of several positive, negative and/or neutral cue words. The AMT has reasonable internal consistency and test-retest reliability, and responses on the AMT typically have a single factor structure related to the number of specific memories retrieved, independent of cue valence (Griffith et al., 2009; Griffith, Kleim, et al., 2012; Griffith, Sumner, et al., 2012; Heron et al., 2012; Takano et al., 2017). It is of note that despite this psychometric consensus, there is little consistency within the literature regarding the use of rAMS or OGM to refer to the presence of impoverished retrieval of specific autobiographical memories or enhanced retrieval of general (categoric/extended) autobiographical memories, with both concepts often being used interchangeably. It is also unclear whether rAMS and OGM are symmetrically opposed to one another such that retrieval of fewer specific memories necessitates enhanced retrieval of general memories, given that it is also possible for participants to give no response at all within the AMT. As such, the present investigation refers to both rAMS and OGM and analyses the presence of both in psychiatric groups.

Nevertheless, this discrepancy between studies within the field, and others like it, mean there is substantial variability between studies regarding their operationalisation of
rAMS/OGM and their reporting of their data. Some studies ignore general memories and either count the total number of specific memories retrieved (Raffard et al., 2010) or the proportion of specific memories relative either to the total number of memories retrieved (Piltan et al., 2020) or the number of cues that were given (Farina et al., 2019). Other studies also (or instead) calculate similar number/proportion scores for both categorical and extended memories together (Kuyken et al., 2006) or just categorical memories (Pollock & Williams, 2001). Also, some studies have modified the parameters of the original AMT, and other variants of cued recall tasks have also emerged. While many studies have used the ‘traditional’ approach of giving participants 60 seconds to retrieve a memory, others provide different durations ranging from 30 seconds up to several minutes, or no specified time-limit (Kleim & Ehlers, 2008).

Similarly, although most studies give five positive and five negative cue words, other studies have used different numbers or types of cue words (e.g., Croll and Bryant (2000) used six positive, negative and neutral cue words). Despite psychometric consensus within AMT studies (Griffith, Sumner, et al., 2012) and findings that impairments in specificity or overgenerality have typically not been found to differ based on cue word valence (e.g., Barry et al., 2018; Liu et al., 2013; Van Vreeswijk & De Wilde, 2004), valence-related differences in specificity/generality remain of interest given the abnormalities in positive and negative valence systems in a range of disorders, including depression, schizophrenia, and PTSD. The original AMT also explicitly instructed participants to retrieve specific memories, but other studies have not given this instruction (Schoofs et al., 2013). Also, most studies have asked participants to respond verbally, where responses are transcribed and then coded, whereas others have asked participants to write or type their responses (Spinhoven et al., 2009).

Clearly, there is tremendous variability in the way that the AMT has been administered and this is to say nothing of the alternative tasks that have been used within this
literature, such as the Autobiographical Memory Interview (Levine et al., 2002). It is important to establish what, if any, effect these variations in methodology have had on the differences in effect sizes that are likely to be observed between studies that have measured specificity/generality in psychiatric groups.

There are also variables related to participants’ characteristics that have been associated with their retrieval of specific/general memories. It remains possible that we might see differences between diagnostic groups in the extent to which they differ from their diagnoses-free counterparts. That is not to say that we hypothesise that the effect size will be substantially larger for particular diagnostic groups compared to others. In fact, existing studies indicate that there is limited difference between clinical groups when they are compared directly, such as when comparing people with Major Depressive Disorder and Bipolar Disorder (Young, Bodurka, et al., 2016) or Acute Stress Disorder (Kleim & Ehlers, 2008). In addition, a review of studies that have explored the underlying neural substrates of autobiographical memory retrieval indicated that, across studies with different clinical populations, there are striking similarities in the network of brain structures that show problematic activity during retrieval of specific memories (Barry, Chiu, et al., 2018). As such, we hypothesise that rAMS/OGM are transdiagnostic constructs and do not anticipate that there would be substantial or consistent differences between diagnostic groups in the extent to which they could retrieve specific or general memories, as compared to controls. However, we do expect that there would be variability between studies in the size of their effects and that these will be explained by other moderator variables.

Another set of participant-related factors that might contribute to rAMS/OGM regard participant age (Barry et al., 2020), gender (Young et al., 2013b), education level (Boelen et al., 2010; Farina et al., 2019; Wessel et al., 2001) and the difference between clinical and control participants in the severity of their symptoms (Liu et al., 2013; Van Vreeswijk & De
Wilde, 2004). In addition, although most studies use interviews with trained clinicians to establish participants’ diagnostic status, some studies use self-report questionnaires with validated cut-off scores for probable diagnoses (Kyung et al., 2016). Again, it is important to establish what effect these differences between participants, and the way in which their diagnoses were determined, have on group differences in effect sizes.

Another set of potential moderators regards the cognitive processes that are theorised to be involved in the retrieval of autobiographical memories. Williams (2006) outlined three processes that explained impoverished retrieval of specific memories and facilitated retrieval of general memories: Capture and Rumination, Functional Avoidance and eXecutive control (CaRFAX). He suggested that, for people with psychiatric problems, specific autobiographical memories are not forgotten, they are just difficult to retrieve. This failure to retrieve occurs when – in accordance with the self-memory system model of autobiographical memory (M. A. Conway & Pleydell-Pearce, 2000) – the search for a memory is aborted too early, such that general-level categorical or extended memories are retrieved instead. This retrieval process could end as a result of capture by and rumination on, or repetitive thinking about, self-relevant information stored at the general level of the memory hierarchy (Crane et al., 2007; Sutherland & Bryant, 2007). It could also be a learnt or functional attempt to avoid the detail and vividness of specific memories and their associated negative affect (Debeer et al., 2014; Hallford, Austin, et al., 2018). The search could also be stopped because one lacks the executive capacity to hold the details of a memory in mind whilst increasingly more event-specific information is retrieved from store, and whilst one also tries to inhibit distraction from our thoughts or things in the environment (Dalgleish et al., 2007).

An evaluative review of existing evidence in this area suggested that although there was some mixed evidence in support of each process, the general consensus within the literature supported the CaRFAX model (Sumner et al., 2014). However, this analysis did not
examine the contribution of these processes to the rAMS/OGM that is observed in psychiatric
groups. In the present study, we report the first meta-analytic examination of how differences
between clinical and control groups in the CaRFAX processes might moderate any
differences between these groups in specificity/generality. To our knowledge, few studies in
this area that include both clinical and control groups report correlation coefficients within
these group separately. As such, we take the approach of calculating standardised effect sizes
for differences between clinical and control groups for each of the CaRFAX processes across
different measurement tools. For each of the CaRFAX processes we are then able to explore
whether, for example, studies that have clinical participants who exhibit a particularly
marked tendency to ruminate, compared to control participants (i.e., a larger effect size) also
show a particularly large effect for differences in the retrieval of specific/general memories.

In summary, we present the first complete meta-analysis of existing studies that test
the difference between people with diagnoses of psychiatric conditions and their diagnoses-
free counterparts in the retrieval of both specific and general autobiographical memories.
Several studies in this area include two or more clinical groups (e.g., a group with depression
and a group with Acute Stress Disorder; Kleim & Ehlers, 2008) and compare these against a
single control group. As such, we use ‘multi-level’ meta-analysis where these study-level
dependencies are accounted for. Our analysis included all psychiatric conditions that had
been studied within the context of rAMS/OGM, with the exception that studies with
participants with Substance Use Disorders were excluded given the potential for these
substances, and not the disorder per se, to influence performance in the study tasks. We
expected that there would be large pooled effects across the included studies in the direction
of impoverished retrieval of specific memories (rAMS) and enhanced retrieval of general
memories (OGM) amongst people with diagnoses compared to controls.
Separate meta-analyses were conducted for group differences in recall for memories cued by positive and negative words. However, given the mixed evidence within the literature, we did not expect that rAMS/OGM effects would differ between cue valence types. In the overall analysis, across cue types, we expected that there would be a substantial amount of heterogeneity between studies in the size of their effects. Meta-regression analyses then examined the contribution of diagnostic (e.g., diagnosis), demographic (e.g., age), methodological (e.g., cue type, number) and process (e.g., CaRFAX variables) moderators to effect size variance. We did not have specific directional hypotheses for most of these variables, except CaRFAX processes for which we expected that higher rumination and avoidance and lower executive functioning scores would be associated with larger effect sizes between psychiatric and control groups.

**Method**

The study procedure was pre-registered within the PROSPERO database [https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019125992](https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019125992)

**Search strategy**

The search strategy involved two iterations. In the first iteration, an online search was conducted using the search engines Proquest and Ovid to search for keywords on the Embase, PsycARTICLES and PsycINFO databases. Keywords included *autobiographical memory* and combinations of either *specificity* or *overgeneral*, different psychiatric diagnoses (e.g. *Depression, Post-Traumatic Stress, Schizophrenia* etc.) and terms related to CaRFAX process (e.g., *rumination, brooding, executive function, verbal fluency, problem-solving*, etc.). A full list of search terms can be viewed in the online supplement. In the second iteration, review articles, including systematic reviews and meta-analysis, were examined for relevant citations that were missed in the first iteration. Also, the authors’ personal collections were reviewed and relevant experts within this field were consulted for additional
American Psychological Association (APA) Meta-Analysis Reporting Standards (MARS) data. The initial search and request for studies was conducted on August 6th 2020. See Figure 1 for a flow chart and the PRISMA checklist for reporting standards is also available in the online supplement.

Two trained research assistants conducted the initial search and subsequent data extraction. An independent research assistant then re-extracted a sample of approximately 20% of datapoints to ensure that the initial coding was accurate. Inter-rater reliability was high (Kappa = .982).

**Study eligibility**

In our pre-registration and in the first iteration of our search, we initially set the criteria for inclusion that studies should be published in or accepted for publication by a peer-reviewed journal. However, in order to bring the meta-analysis in line with American Psychological Association (APA) Meta-Analysis Reporting Standards (MARS) we also circulated a message to an Autobiographical Memory and Psychopathology email list to ask if colleagues (N = 101) had any unpublished data that could be made available. No unpublished data were reported by colleagues and so no additional data, besides those reported in published manuscripts, were included. No upper limit on the age of a study was set.

To reiterate, the majority of studies within this field use the Autobiographical Memory Test (AMT; Williams & Broadbent, 1986). The AMT typically gives participants 60 seconds to recall autobiographical memories in response to each of several positive, negative and/or neutral cue words, though there is substantial variability in the parameters of the AMT and the operationalisation of specificity/generality therein. In addition, not all studies use the AMT. As such, studies did not have to use the AMT in order to be included and we did not specify that studies must report specific versus non-specific memory scores, or cue-specific or total scores.
Instead, inclusion was determined by the measurement – and availability of mean and standard deviation data for – either the number or proportion (either as a function of the total number of memories asked for, or as a function of the total number of memories given by participants) of specific or general memories. We included studies that report these scores across cue types or for each cue type separately. Several studies (e.g., Corcoran & Frith, 2003; Cuervo-Lombard et al., 2012; Danion et al., 2005; Kaviani et al., 2005; Potheegadoo et al., 2012) were excluded because they presented a mean score from a continuous scale from general memories to specific which made it impossible to separate the frequency of specific and general memories retrieved by each group of participants.

Memories could be reported in a written, typed, or oral format and the memories had to be reported immediately after the cue word was given. Several neuroimaging studies (e.g., (Parlar et al., 2018; Young, Erickson, & Drevets, 2012; Young, Erickson, Nugent, et al., 2012) were excluded because in these studies participants were instructed to recall their memories, focus their thoughts on them for a period of time whilst in the scanner, then complete a distractor task, and finally report the memories for coding once outside of the scanner. This distractor task introduced a dissociation between the retrieval and reporting phases of the experiment. If studies included a within-participant manipulation where participants were constrained to recall memories related to specific themes (e.g., trauma or cancer (Harvey et al., 1998; Kangas et al., 2005) and a condition without such constraint, data were recorded from the unconstrained condition. Also, in the event that a study reported longitudinal data at several timepoints, data were taken from the earliest timepoint (Kangas et al., 2005). In the event that a study included two forms of AMT, such as one with cue words and one with images (Ridout et al., 2016), data from the cue word task were extracted given the commonality of this procedure within the literature. These decisions were made to reduce heterogeneity between studies where the number of studies that included such modifications
(e.g., asking for particular memory themes) was too small for us to analyze these changes as moderators of effect sizes.

To maintain consistency between the included studies, it was also a requirement that recalled memories were coded in terms of specificity/overgenerality by a trained researcher. This excluded studies in which participants coded their own memories (St. Jacques et al., 2013; Young, Drevets, et al., 2016) or only pressed a button to indicate that a specific memory had been retrieved but did not report the actual memory (Hach et al., 2014).

Studies must have compared a group of participants who had been assigned a psychiatric diagnosis with a group of people who were diagnoses-free. Classification of diagnoses had to be conducted using the criteria set out in either the Diagnostic and Statistical manual (DSM) or International Classification of Diseases (ICD) and could be classified either through interview with a trained clinician or through the use of self-report questionnaires that had been clinically validated. Given the breadth of years that the sampled studies came from, no specific version of these diagnostic systems was referred to. The only diagnostic group that was excluded from the study was Substance Use Disorders given the potential for these substances, and not the disorder, to influence performance in the study tasks.

In the event that a study included two control groups (e.g., Piltan et al., 2020; Wittekind et al., 2017), the control group that best represented a healthy community sample was selected (i.e., one that was free from adversity/trauma and diagnoses). Studies were excluded if they involved participants that were in remission from their primary diagnoses or where the control group was characterized as having participants that reported that they had previously received diagnoses, with one exception: Park et al. (2004) included participants who were reported as having full current diagnoses of MDD and those that were reported as being in partial remission. Data were presented collapsed across these groups. Given the way
the data were presented and the fact that it was also reported that both groups nonetheless fulfilled DSM criteria for MDD, this study was included, and these combined data were extracted. The remaining studies that categorized participants as being in remission were excluded on the basis that such participants were unlikely to pass a diagnostic threshold and that an analysis of the extent to which rAMS/OGM is present amongst people in remission warrants a separate, dedicated analysis.

Studies could be included if they involved an intervention but where there was a pre-treatment assessment of clinical-control differences or if participants in the clinical and control groups were both given a placebo (Wingenfeld et al., 2012, 2013). Whether participants were under placebo medication during the measurement of memory specificity/generality was recorded.

At the end of the search process, the majority of exclusions were because studies did not include a measure of specificity/generality or because their operationalisation of these variables was not appropriate for inclusion ($n = 116$; e.g., they used a scale for degrees of specificity, Corcoran & Frith, 2003), because they did not include a diagnoses-free control group ($n = 86$) or because the data could not be sourced, even after contacting the authors ($n = 31$). See Figure 1 for further information.

**Data extraction and handling**

*Participant characteristics*

We extracted the number of participants in each group within a given study, their mean age, the number of females and the average number of schooling years completed by the clinical group. The primary diagnosis of the clinical group was recorded and whether the diagnosis was provided by a clinician or a clinically validated questionnaire. Where a study explicitly reported the ethnicity of participants, this was also extracted.
As there were a large number of diagnoses, where there were fewer than three studies including a particular diagnosis, we grouped these with other diagnoses so that the data could be analysed with greater power. This process resulted in the following diagnostic categories: Mood Disorders (Major Depressive Disorder and Bipolar Disorders), trauma-related disorders (PTSD and Acute Stress Disorder), Psychosis (Schizophrenic disorders), Anxiety Disorders (including Obsessive Compulsive Disorder), Borderline Personality Disorder, and other disorders (Eating disorders, complicated grief, studies with a mixture of diagnoses, avoidant and dependent personality disorders and somatic symptom disorder). Where available, the proportion of participants taking medication, the average age of illness onset (years), and the average illness duration (months) were recorded.

With some notable exceptions, we also extracted the mean and standard deviation scores for both the clinical and control group for the primary measure of symptom severity used within each study. Several studies included more than one depression measure. In all cases the Beck Depression Inventory (BDI) was selected over the other measure, given the prevalence of the BDI within the rest of the studies that were sampled. A small number of studies only recorded the symptoms of the primary diagnosis in their clinical group but measured depression symptoms in both groups (Ricarte et al., 2014; Wittekind et al., 2017) or they only measured depression symptoms even though depression was not the primary diagnosis of the participants in the study (Barry, Del Rey, et al., 2018). In these cases, the depression measure was included as the measure of symptom severity even though the primary diagnosis category for the clinical groups in these studies remained unchanged (i.e., Schizophrenia or PTSD). The standardized mean difference for severity scores between groups for each study was computed and this variable was used as a predictor within meta-regression models.

*Specific and general autobiographical memories*
We extracted the mean and standard deviation, for both groups, for the number or proportion of specific and general memories retrieved across cue types and for each cue valence separately. We codified studies based on whether or not (and how) the task measured specific or general memories, whether the instructions of this task explicitly specified to participants that they should retrieve specific memories or not, the duration of time participants were given to recall each memory following cue presentation, whether responses could be given verbally or otherwise, and the number of cues given to participants.

Throughout the analyses, our intention was to test between-group differences within each cue valence and across cue types. To accommodate differences between studies in the reporting of data, and to maximize the number of studies that could be included within each analysis, some data imputation was necessary. In particular, as a minority of studies only reported cue-specific scores and not total scores, total scores were imputed by averaging the mean and standard deviations across cue valences. This enabled us compute effect sizes for these studies within the test of between-group differences in specificity/generality across cue types, whilst also testing between-group differences within each cue type.

Similarly, some studies that quantified general memories gave scores for both categoric and extended memories separately and others gave an aggregate score across these memory types. We did not have any hypotheses about between-group differences within categorical and extended memories. As such, where a study gave scores for only one of these memory types (e.g., only categoric memories), this was used within our overall analysis across cue types. Where a study reported both categorical and extended memories and did not report an overall general memory score, an average of these scores was computed. Again, this enabled us to include these studies in the computation of effect sizes for the overall test of between-group differences.
One study reported both proportion scores and raw means. In this case, as the raw means gave the most complete picture of retrieval, the raw means were retained (Smets et al., 2014).

Process variables

Mean and standard deviations were also extracted for both groups for measures of repetitive negative thinking (measures of rumination, or its sub scores of brooding or reflection, and measures of worry), cognitive avoidance, and executive functioning (working memory and verbal fluency). One study included two measures of rumination (Smets et al., 2014) but as one of these, the Ruminative Response Scale (RRS) was more commonly used by other studies sampled, scores for this measure were retained. One study included the RRS but only reported sub-scores and not the total. As the brooding subdimension of the RRS better captures the unconstructive form of rumination that is thought to impact retrieval of specific memories (Williams et al., 2007), scores for the brooding subdomain were included rather than averaging the two scores. With regards to avoidance, one study (Ganly et al., 2017) included three scales: the White Bear Suppression Index, the Cognitive-Behavioral Avoidance Scale (CBAS) and the Acceptance and Action Questionnaire. Because the CBAS is conceptually closer to the other kinds of self-report avoidance measures that were included (e.g., Cognitive Avoidance Questionnaire), this was retained. Finally, where a study included two or more indices of working memory (e.g., forward and backward digit span; Schönfeld et al., 2007) or verbal fluency (e.g., semantic and phonemic; D’Argembeau et al., 2008) scores on these measures were averaged. As with disorder severity, the standardized mean difference between groups for each study was computed for each process variable and these mean difference scores were used as predictors within meta-regression models.

Risk of bias assessment
Several possible forms of bias were recorded during data extraction. First, we coded whether the study involved a randomization between and within the study tasks (e.g., were cue words presented in a fixed or random order). Second, we noted whether participants’ group allocation was concealed from them. Third, we noted whether participants and personnel were blind to the nature of the study. Fourth, we noted whether the coders for the autobiographical memory task were blind to participants’ group designation and the nature of the study during coding. Fifth, we noted whether there was incomplete outcome reporting or whether particular participants were included in the study but were omitted from analyses for unclear reasons. Finally, we noted whether scores for particular measures were included in the study but were selectively omitted from the final report. Where a study showed explicit evidence of any of these biases, a high degree of bias was noted; where the study did not include enough information for us to assess clearly whether the bias was present or not, some concern was noted. Where the study explicitly noted the steps taken to manage the bias mentioned, we noted that there was a low degree of bias.

Analysis strategy

To account for dependence between effect sizes (same control group), three-level random effects univariate meta-analyses were conducted with maximum likelihood estimators using the metaSEM package (Cheung, 2014), where effect sizes for clinical-control comparisons were nested within studies. Effect size heterogeneity was reported in terms of $I^2$, $r^2$ and $Q$. These analyses were conducted for effect sizes across cue types and for each cue valence separately. Where there was evidence of a significant difference between pooled effect sizes for each cue type then subsequent moderator analyses were planned to be conducted for each cue valence separately. Otherwise, moderator analyses focused on across-cue-type effect size variance.
This moderator analysis examined the extent to which differences in method, demographics and process-related variables between studies accounted for heterogeneity in effect sizes for specific/general memory retrieval. Overall, the role of 15 moderators was explored. Six variables related to participant characteristics: 1) whether the diagnosis was given by clinician interview or a validated questionnaire measure; 2) participant’s diagnostic category; 3) mean age of participants; 4) the proportion of females in the study; 5) average number of years in education; and, 6) differences in symptom severity between clinical and control participants. A further five variables related to the measurement of specificity/generality: 1) the instruction given to participants prior to retrieval (traditional AMT vs. other); 2) whether participants could respond orally or written; 3) the manner in which specificity/generality was handled (number vs. proportion); 4) the number of cues given to participants; and, 5) duration of time given for retrieval. Finally, four more variables related to processes involved in memory retrieval: 1) rumination; 2) avoidance; 3) working memory/executive functioning; and, 4) verbal fluency.

Given the number of potential moderators that were sampled, multimodel inference was used to find the most important predictors of effect sizes and so reduce the number of variables that would subsequently be included in three-level metaregression models. However, the dmetar package (Harrer et al., 2019) and multimodel inference have some limitations that are of note. First, multimodel inference within dmetar excludes studies in a listwise manner if they are missing data for any included predictors. As such, multimodel inference analyses only included participant characteristic and methodological variables as these were available in most studies. Process variables (rumination, working memory, avoidance, verbal fluency) were only measured in a relatively small subset of studies and so these variables were included in a separate set of analyses not using multimodel inference.
Second, there is not yet any way to conduct multimodel inference in a three-level meta-analysis model. As a compromise, multimodel inference was performed on the basic two-level model (only accounting for variance between effect sizes and not study-level dependence) such that important variables could then be selected and tested within a subsequent three-level metaregression. Separate multimodel inference models were conducted for specific memory retrieval and another for general memory effect sizes. The same moderators were input into each of these two models and those moderators whose model-averaged importance exceeded .8 (that is, they were included in the majority of models with good fit) were designated as important (Harrer et al., 2019).

Important variables were then included in subsequent three-level metaregressions predicting each of specific and overgeneral memory effect sizes to examine if their contribution to effect sizes continued even when study-level dependencies were accounted for. As the process variables were only measured in a small subset of studies, many of which only measured one of each kind of process variable, these variables were each entered into separate three-level metaregression models to examine their contribution to effect sizes. These models replicated the previous three-level models using metaSEM package (Cheung, 2014) with maximum likelihood estimators.

There are limited methods currently available for assessing publication bias within three-level meta-analyses. Following the guidance of Harrer et al. (2019) we present funnel plots and a rank correlation test (Kendall’s τ) that examines publication bias in terms of whether larger effects are associated with smaller sample sizes. As further tests of selective reporting, we also conduct Trim and Fill and P-curve analyses using two-level models that do not account for study-level dependencies. In addition, to test for the contribution of potential sources of bias (e.g., blinding) to between-group effect sizes, we conducted two more multimodel inference models predicting between-group differences in retrieval of specific
and general memories. For these models, the risk of bias scores which are otherwise coded as low risk, some concern of risk and high risk, were recoded such that some or high risk of bias were coded as “1” and low risk was coded as “0”. This was because the some concern of risk code was typically given when there was no clear information about the potential risk of bias and so the decision was taken to treat these studies similar as high risk studies for the purposes of these analyses.

Results

Study characteristics

See Table 1 for a summary of study characteristics separated by analyses for memory specificity and overgenerality. Across these analyses, the sampled studies (N = 74) included 98 effect sizes for comparisons between different clinical groups and a control group. 85 of these comparisons were for differences in autobiographical memory specificity across cue types, within which there were 43 effect sizes for differences in recall of specific memories related to positive and negative cues separately. The studies from which these effects were extracted had a median publication year of 2009 (range: 1996-2020). 56 comparisons were for differences in the recall of general memories across cue types, 28 of which were for positive and negative cues separately. The studies from which these effects were extracted had a median publication year of 2010 (range: 1998-2020).

Participant characteristics

Across these comparisons there were 7536 participants (nclinical = 3122; ncontrol = 4414). The grand mean age across studies was 34.70 years (SD = 12.37; Clinical: Mage = 35.62, SDage = 12.48; Control: Mage = 33.80, SDage = 12.73; k = 7 did not provide age data). The average percentage of females across the sampled studies was 63.4% (Femalesclinical = 1802 (57.7% of clinical participants); Femalescontrol = 1906 (43.2% of control participants; k = 5 did not provide gender data). Participants in the clinical groups received an average of 10.49 years of
education ($SD = 3.41; k = 34$). Within the few studies that explicitly referred to their sample ethnicity, participants were overwhelmingly white ($Mean proportion = .71; SD = .24; k = 12$).

Forty-four effect sizes included participants with Mood Disorders (Bipolar Disorder: $k = 2$; Major Depressive Disorder $k = 42$), 14 included participants with psychosis (Schizophrenia $k = 11$; Delusional Disorder $k = 1$), 12 included participants with trauma-related disorders (Post-traumatic Stress Disorder $k = 9$; Acute Stress Disorder $k = 3$), 9 included participants with Borderline Personality Disorder, 7 included participants with Anxiety-related disorders (Specific Phobia $k = 1$; Social Anxiety Disorder $k = 1$; Mixed Anxiety $k = 3$; Obsessive Compulsive Disorder $k = 1$; Obsessive Compulsive Personality Disorder $k = 1$) and 14 were included within an other category (Mixed diagnoses $k = 4$; Mixed-suicidal $k = 2$; Complicated Grief $k = 2$; Anorexia $k = 1$; Avoidant Personality Disorder $k = 1$; Dependent Personality Disorder $k = 1$; Mixed Eating Disorders $k = 1$; Externalizing Disorder $k = 1$; Somatic Symptom Disorder $k = 1$). These diagnostic categories were created to closely map onto the diagnostic groupings that exist within DSM. These groupings were then used to examine differences between diagnoses in our moderator analyses whilst limiting the number of predictors that would need to be included within our models relative to if individual diagnoses were used as predictors. Unfortunately, this left us with a miscellaneous other category given the limited number of studies that were sampled in some diagnostic groups.

93 of these diagnoses were clinician-determined and 4 were questionnaire-determined. Within the few effects that had reported data on medication use ($k = 20$) the majority of participants were receiving medication ($Mean proportion = .77; SD = .32$). There were data on illness duration for 13 effects ($Mean months = 151.08; SD = 92.08$) and 7 had associated data on illness onset ($Mean age = 21.31; SD = 1.21$).

Cuing procedure
**Autobiographical memory task.** 92 effect sizes used the Autobiographical Memory Test (AMT) or some variant of it. Of these, 77 used a traditional AMT where participants were asked to retrieve specific autobiographical memories in response to a number of adjectives/adverbs (e.g., happy); 8 used a version of the AMT wherein they were cued to think of specific instances in the past where they exhibited particular traits (e.g., jealousy) (Kremers et al., 2004, 2006; Spinhoven et al., 2009); 3 involved a variant on the AMT where participants were instructed to retrieve specific autobiographical memories in response to cue words but they were also asked to ‘think aloud’ their entire thought process during retrieval and not just the memory that was cued (Barnhofer et al., 2002; Heidenreich et al., 2007); 2 used an alternating instructions AMT where participants were instructed to switch between recalling specific and general memories (Hitchcock, Rodrigues, et al., 2019; Piltan et al., 2020), 2 used a minimal instructions AMT where they were not instructed to retrieve memories of any particular specificity/generality (Schoofs et al., 2013). Of the remaining effects, 1 was obtained from a study using the Twenty Statements Task where participants were asked to give 20 descriptions of themselves by completing sentence stems (“I am…”) with reference to past autobiographical events (Bennouna-Greene et al., 2012), 1 included a paradigm where participants had to think of multiple good and bad days from their past (Hitchcock et al., 2020), 1 included the self-defining memory questionnaire which specifically cues self-defining memories (Raffard et al., 2010) and 3 studies used each of the Autobiographical Memory Interview (Mehl et al., 2010), the Autobiographical Memory Enquiry (Nieto et al., 2018), and the Bielefelder Autobiographical Memory Inventory (Oertel-Knöchel et al., 2012), each of which cue participants to think about life periods. Across these tasks, 80 effects came from studies that examined verbal responses to the memory tasks and 12 came from studies that involved written/typed responses ($k = 6$ did not report how participants gave their responses).
**Cues.** Most effects \((k = 50)\) came from studies that gave participants 10 cues. Also, the median number of cues given was 10 with a range of 2-36. In particular, 1 effect came from a study that used 2 cues (Good days and bad days; Hitchcock et al., 2020) and 1 effect came from a study that used 36 cues (12 positive, 12 negative, 12 neutral; Maurex et al., 2010). Of the effects that came from studies where participants were given valenced cues, 78 included positive and negative cues, 12 included positive, negative and neutral cues, 2 included only positive cues (Schoofs et al., 2013), and 1 included positive and negative cues and also trauma-related cues (Wittekind et al., 2017). Of the remaining 5 effects, 3 cued life periods (Mehl et al., 2010; Nieto et al., 2018; Oertel-Knöchel et al., 2012) and 2 used self-defining memory cuing tasks (Bennouna-Greene et al., 2012; Raffard et al., 2010).

**Recall duration.** Most effects came from studies that gave participants 60 seconds to retrieve a memory \((k = 54)\). Of the remaining, 20 gave participants 30 seconds, 7 gave participants 120 seconds, 1 gave participants 240 seconds, and 6 gave participants unlimited time. 10 studies did not specify the amount of time that participants were given.

**Operationalization of specificity/overgenerality.** Of the 85 effects that quantified specific memories, 57 operationalized this as the number of specific memories retrieved, 9 operationalized this as the proportion of specific memories relative to the number of memories given (total cues minus the number of omissions), 5 operationalized this as the proportion of specific memories relative to the total number of cues given, and 14 operationalized this as the proportion of memories given but did not specify whether this was relative to memories or cues given.

Of the 56 effects that quantified general memories, 18 of these considered only categorical memories, 35 considered both categorical and extended memories and 1 did not specify what was meant by general memories (Iqbal et al., 2004). Across these, 46 used the number of general memories retrieved, 5 considered the proportion relative to memories
given, 2 considered the proportion relative to cues given and 3 did not specify what their proportion was relative to.

**Process variables.** A minority of studies measured CaRFAX variables. Of the 14 effect sizes that included a measure of rumination, 10 of these used the Ruminative Response Scale (RRS). Of the 11 effect sizes that included a measure of avoidance, 6 used the Acceptance and Action Questionnaire (AAQ). For executive functioning, eleven studies included a measure of verbal fluency. The majority of these additionally included another measure of executive functioning so separate variables for verbal fluency and other measures of executive functioning were created. Of the remaining 13 effects included in this latter variable, 12 included measures that are typically considered to be measures of working memory (e.g., a test of digit span) and one effect included Part B of the Trail Making Test that measures inhibitory control (Reitan & Wolfson, 1993).

**Specific memory retrieval**

**Between-group differences**

Clinical and control groups showed a large difference in their overall autobiographical memory specificity and this effect differed significantly from zero \((k = 85; g = -0.864, 95\% CI[-1.030, -0.698], Z = -10.223, p < 0.001)\). Similarly sized effects were also evident for group differences in autobiographical memory specificity following positive cues \((k = 43; g = -1.007, 95\% CI[-1.189, -0.826], Z = -10.891, p < 0.001)\) and negative cues \((k = 43; g = -0.719, 95\% CI[-0.998, -0.441], Z = -5.063, p < 0.001)\). That these effects differed significantly from zero suggests that clinical groups have greater difficulty retrieving specific autobiographical memories across cue types and for positive and negative cues. However, the overlap in the confidence intervals for the effects for positive and negative cues indicates that clinical groups are unlikely to differ from control groups in their ability to retrieve specific
memories for positive cues, compared to negative cues. See Figure 2 for a forest plot and Table 2 for a breakdown of pooled effect sizes by diagnostic group.

**Moderator analyses**

Given the absence of the valence effect and the substantially smaller number of available studies within these analyses ($k = 43$) compared to the overall across cue-type analysis ($k = 85$), subsequent moderator analyses focused on examining the factors that explain variance in effect sizes for specific memory retrieval across cue types. In this model, the total variance not attributable to sampling error ($\hat{\tau}^2$) was $.850 (Q(84) = 390.148, p < .001; Effect-level, $\hat{\tau}^2 = .188, \tau^2 = .086; Study-level, \hat{\tau}^2 = .662, \tau^2 = .301$).

**Multimodel inference.** Due to the limitations of the R package being used, potential study and demographic moderators were tested using two-level multimodel inference (see Figure S1 in the online supplement for full list of variables and importance scores). This method tests multiple combinations of predictors to determine which combinations of predictors provide the best explanation of effect size variability; it assigns importance values to each predictor equal to the sum of the weights/probabilities for the models in which the predictor appears. Those predictors that are present in most of the best fitting models – or which have importance coefficients greater than .8 – are described as *important*. Across the models that were fitted, a high degree of predictor importance was obtained for the degree of difference between clinical and control groups in their symptom severity (1.00), duration (1.00), the mean age of participants (.96) and the proportion of female participants in each group (.90). As the duration variable was a categorical variable, the multimodel inference model also gave separate coefficients for dummy variables for each level duration. By doing this, the variable that corresponded to whether the retrieval duration was 30 seconds or longer showed a significant effect (Estimate = -1.795, $SE = 0.656, Z = 2.738, P = 0.006$). None of the other dummy variables for other levels of duration showed significant effects (smallest $P = .158$).
Symptom severity, age, gender and the 30 second duration dummy variable were therefore retained for subsequent moderator analyses within the three-level model.

The remaining variables (participants’ diagnosis category (depression, trauma-related, psychosis-related, anxiety-related, Borderline Personality Disorder vs. other), the way in which diagnoses were assigned (clinician vs. questionnaire-cutoff), the instruction given to participants prior to retrieval (traditional AMT vs. other), the manner in which specificity was handled (number vs. proportion), the number of cues given to participants, the response format (verbal vs. written)) showed low levels of predictor importance (highest value was how specificity was handled, .61). See Figure S2 for full presentation of predictor importance scores.

**Three-level moderator analysis.** In the three-level model with multiple predictors of effect size examined simultaneously, there was a significant effect of duration, such that larger, more negative, differences between clinical and control groups were evident in studies where participants were given only 30 seconds to retrieve memories compared to longer durations, \( b = -0.784, SE = 0.225, 95\% CI[-1.226, -0.342], Z = -3.482, P < .001 \). There was a marginal effect of participants’ age such that larger differences between clinical groups and controls were evident in studies with participants who were younger, \( b = -0.013, SE = 0.007, 95\% CI[-0.026, -0.001], Z = -1.882, P = 0.059 \). There were no significant effects for participants’ symptom severity, \( b = -0.026, SE = 0.035, 95\% CI[-0.095, 0.042], Z = -0.757, P = 0.449 \), or the proportion of females in the study samples, \( b = 0.042, SE = 0.313, 95\% CI[-0.572, 0.656], Z = 0.135, P = 0.893 \).

Although we planned to include process variables in these moderator analyses, there were so few studies that included these variables that there were errors in fitting these models when other moderators were also included. As such, separate models for each of the process variables were conducted. In each of these models, none of the process variables reached
statistical significance in explaining variance in effect sizes, although the finding was marginal for working memory/executive functioning, whereby larger group differences were associated with larger effects on specificity ($b = 0.925, SE = 0.530, 95\% CI[-0.114, 1.963], Z = 1.746, P = 0.081$). Lastly, the average numbers of years that participants in the clinical group spent in education was also not associated with study variance, $b = -0.061, SE = 0.039, 95\% CI[-0.137, 0.0158], Z = -1.554, P = 0.120$. See Table S1 for full results.

Pooled effect sizes indicated that people with psychiatric diagnoses had substantial difficulty retrieving specific autobiographical memories, relative to diagnoses-free control groups, and this was true for retrieval following both positive and negative cues. However, there was also a large amount variability between studies in the size of these effects. There was evidence that studies with short, 30s, retrieval durations had larger differences between psychiatric groups and controls. There were no other significant moderators of effect sizes.

**General memory retrieval**

*Between-group differences*

Consistent with findings on specificity, clinical groups and control groups also showed a large difference in their retrieval of general autobiographical memories and this effect differed significantly from zero ($k = 56; g = .712, 95\% CI[0.524, 0.900], Z = 7.437, p < 0.001$). Similarly sized effects were also evident for group differences in general memory retrieval following positive cues ($k = 28; g = 0.715, 95\% CI[0.511; 0.919], Z = 6.879, p < 0.001$) and negative cues ($k = 28; g = 0.725, 95\% CI[0.523; 0.927], Z = 7.031, p < 0.001$). Clinical groups recalled more general memories than control groups overall and these effects were similarly sized for memories retrieved following positive and negative cues. As in the analysis of specific memory retrieval, the overlap in effect sizes for positive and negative cues indicated that there was little evidence of cue-specific effects for general memory
retrieval. See Figure 3 for a forest plot and Table 2 for a breakdown of pooled effect sizes by diagnostic group.

**Moderator analyses**

Similarly, there were substantially fewer studies in the valence-specific analysis ($k = 28$) than the overall across cue-type analysis ($k = 56$). As such, subsequent moderator analyses focused on examining the factors that explain variance in effect sizes for general memory retrieval across cue types. In this model, overall $I^2$ was .851 ($Q(55) = 294.899, p < .001$; Effect-level, $I^2 = .640, \tau^2 = .287$; Study-level, $I^2 = .211, \tau^2 = .094$).

**Multimodel inference.** Repeating the two-level model with study and demographic variables included as moderators using multimodel inference indicated that group differences in the severity of symptoms (1.00), whether general memory was operationalised using the total number of general memories or the proportion of memories (.96), and participants’ mean age (.92) were each predictors of importance. The next most important variable was the proportion of females (.68)(See Figure S2 for full presentation of predictor importance scores). However, in the multi-level model with these important moderators of effect size examined simultaneously, no predictors had significant effects (smallest $p = .937$ for the way in which general memories were operationalised; See Table S2 for full results).

**Three-level moderator analysis.** As in the analysis of specific memory retrieval, none of the process variables explained a significant amount of variance in effect sizes in their separate models (the strongest predictor was standardised group mean differences for verbal fluency, $b = -0.863, SE = 0.535, 95\% \text{ CI}[-1.913, 0.186], Z = -1.614, P = 0.107$). However, average number of school years did explain a significant amount of variance in effect sizes ($b = 0.091, SE = 0.036, 95\% \text{ CI}[0.021, 0.162], Z = 2.559, P = 0.011$), with more years of education in the psychiatric groups predicting larger differences in overgeneral memories between psychiatric and control groups. See Table S3 for full results.
Pooled effect sizes indicated that people with psychiatric diagnoses recalled more general autobiographical memories, relative to diagnoses-free control groups, and this was true for retrieval following both positive and negative cues. There was also a large amount variability between studies in the size of these effects. In this analysis, studies that included psychiatric groups with more years in education showed larger differences between psychiatric and control groups in the retrieval of general memories. There were no other significant moderators of effect sizes.

**Sources of bias**

*Publication bias*

There was evidence of significant asymmetry in the funnel plots for effect sizes for specific and general memory retrieval and their associated sample variances (see Figure 4). This was supported by a rank correlation test (Specific memory: Kendall’s τ = -0.275, p < .001; General memory: Kendall’s τ = 0.314, p < .001) indicating that studies with larger effect sizes had smaller samples.

As multi-level model packages in R do not offer any other tests of publication bias, we removed the third level and computed two-level models for specific and general memory retrieval using metagen() within the *meta* package in R. In the analysis of specific memory retrieval first, removal of outliers (effects that did not overlap with the pooled effect) still yielded a significant pooled effect size \( (k = 41, g = -0.912, 95\% \text{ CI}[-0.964; -0.860], Z = -34.46, p < 0.001) \). Trim and Fill analyses indicated that correcting for funnel plot asymmetry by replacing 20 effects still yielded a significant pooled effect size \( (g = -0.570, 95\% \text{ CI}[-0.770, -0.370], Z = -5.60, p = < 0.001) \). Similar findings were evident in our analysis of general memory retrieval, with a significant pooled effect size after removing outliers \( (k = 32, g = 0.719, 95\% \text{ CI}[0.634, 0.803], Z = 17.29, p < 0.001) \). There was also a significant
pooled effect present within the Trim and Fill analysis after replacing 16 effects ($g = 0.388$, $95\%$ CI$[0.155, 0.621]$, $Z = 3.32$, $p = 0.001$). P-curve analysis on both sets of data indicated that the data for specificity and generality both contained evidential value (all tests of skewness, $p < .001$) and that evidential value was not absent or inadequate (all tests of flatness, $p > .999$). Put otherwise, our findings were unlikely spurious or due to selective reporting.

**Risk of bias**

As Figure 5 illustrates, there was concern regarding possible sources of bias in each of the assessed bias domains. The majority of studies (weighted percentage: $>75\%$) reported that their tasks were given in a fixed order, and few studies reported whether group allocation was concealed ($>75\%$), whether participants and personnel were blind to the nature of the study ($>60\%$) or whether coders were blind to participants’ group allocation or the nature of the study during outcome assessment ($>75\%$). Due to the considerable lack of pre-registration, it was also unclear whether there was complete data reporting or whether particular measures were selectively omitted from analyses.

Two subsequent multimodel inference models tested whether the presence of, or concerns about, bias were associated with specific and general memory retrieval. None of the possible sources of bias was significantly associated with between-group differences in recall of either of these memory types and all importance coefficients were less than .80 (see Table S4 for a full outline of model estimates and importance coefficients).

**Discussion**

This meta-analysis examined whether, across studies of people with various psychiatric disorders, clinical groups differed from control groups in their ability to retrieve specific and general autobiographical memories. We also examined whether the size of these differences was explained by differences in methodology or the demographic, clinical or cognitive
characteristics of participants in the studies. We found that, amongst the studies sampled, people with psychiatric diagnoses typically recalled significantly fewer specific autobiographical memories and more general autobiographical memories than people without diagnoses. Although the overall effects were large in size, there was a substantial amount of variability between studies regarding their effect sizes. Within our analysis of the factors that explained this variability, three findings were particularly striking. First, as hypothesised, there was no significant difference in memory retrieval based on which diagnostic category participants were in. Second, among the other possible moderating factors, participants with diagnoses recalled even fewer specific memories than controls when they had less time (30 seconds) to recall their memory, compared to longer durations. Third, and contrary to our hypotheses, there was little-to-no clear evidence that group differences in (CaRFAX) processes thought to underlie memory retrieval contributed to variability in group differences in memory retrieval.

The findings of our overall multi-level meta-analysis, and the finding that participants’ diagnoses were not a significant moderator of effect size variance, to some extent support the transdiagnostic nature of autobiographical memory specificity/generality. These possible transdiagnostic effects were further supported by the pooled effects within each diagnostic group (see Table 2), with the possible exception of the anxiety disorder category where the pooled effect size did not differ significantly from zero. Overall, however, the size of the pooled effects, across diagnostic groups, were in the .7 to .9 range, suggesting that difficulty retrieving specific memories and a tendency to retrieve general memories, are significant and pervasive across a range of psychiatric groups. It is striking, therefore, that these problems are not targeted in traditional cognitive-behavioural therapies and that they may even interfere with responsiveness to these interventions (Harvey et al., 2014; Hitchcock, Rudokaite, et al., 2019). The findings presented here therefore support the need
for interventions such as Memory Specificity Training (MeST) in a range of different diagnostic groups (Barry et al., 2019; Hallford et al., 2020; Raes et al., 2009) and support the need for future research that further refines MeST given suggestions that its effects may only be transitory (Barry et al., 2019).

However, it must be noted that the majority of studies sampled involved participants with Major Depressive Disorder (~49%), Schizophrenia (~19%) and Posttraumatic Stress Disorder (~17%) with few studies involving other groups of participants (see Table 1 and Table 2 for a breakdown by analysis). Although our data indicate that rAMS/OGM is reliably shown amongst people with these diagnoses, we cannot make similar claims for all psychiatric diagnoses as many of them were not represented within the sampled studies. It is unlikely that investigations in other diagnostic groups have taken place but have not been published due to non-significant findings, as our search process involved contacting colleagues about their unpublished data and no additional data were offered. Future research in this area must test whether rAMS/OGM is present in a more diverse area of diagnostic groups for the transdiagnosticity of rAMS/OGM to be fully tested. To facilitate this endeavour, and to enable the efficient meta-analysis of future investigations including more diverse sample groups, the data extracted in this study are fully available online for other researchers to add to once more data are available.

One group of diagnoses for which there are remarkably few studies within the literature is Anxiety Disorders (e.g., specific phobia, social anxiety disorder, generalized anxiety disorder). Although Anxiety Disorders are among the most prevalent diagnoses that exist (Kessler et al., 2005) our analysis included only two studies with mixed anxiety samples (Rawal & Rice, 2012; Wessel et al., 2001), one with specific phobia (Kleim & Ehlers, 2008) and one with social anxiety disorder (Heidenreich et al., 2007). Two other studies included groups of participants with diagnoses whose inclusion within the Anxiety-related Disorders
category is debated: Obsessive Compulsive Disorder (Wilhelm et al., 1997), and Obsessive Compulsive Personality Disorder (Spinhoven et al., 2009). Some studies within this grouping were excluded because the necessary data were not available (e.g., Burke & Mathews, 1992). It is of note that, among these included studies, the within-group effect size for the Anxiety Disorder diagnostic category was not statistically significant. This finding contrasts with studies in community samples that indicate that trait and state levels of general anxiety are correlated with rAMS (Hallford et al., 2019; Hallford, Noory, et al., 2018; Hallford & Mellor, 2017). Our finding may be explained by the small sample size and heterogeneity between diagnoses included within this category.

Aside from these diagnostic limitations, our analysis indicates that consistent with psychometric evidence (Griffith et al., 2009; Griffith, Kleim, et al., 2012; Griffith, Sumner, et al., 2012; Heron et al., 2012; Takano et al., 2017), problems with autobiographical memory specificity are unlikely to differ as a function of cue valence. It is important to note that an analysis of cue valence is not equivalent to an analysis of memory valence. Although evidence indicates that there is often congruence between cue and memory valence (Johnstone Ganly, 2017; Nelis et al., 2013; Young, Erickson, & Drevets, 2012), it is possible that a person may see a positive cue (e.g., happy) and retrieve a memory of the opposing valence (e.g., a time when they were not happy). Indeed, MeST includes a session that focuses on these kinds of generalisations (Raes et al., 2009). Previous studies that have coded memory valence have found that people with MDD recall fewer specific positive memories than healthy participants (Young et al., 2013a, 2014, 2015) and that people with schizophrenia diagnoses have particular difficulty recalling negative specific memories (Barry, Del Rey, et al., 2018). As such, it remains possible that there are valence-specific effects when the valence of memories are measured. Future investigations should code memory valence (Ricarte et al., 2011, 2015) in order to inform the development of
interventions such as MeST and the focus that they attribute to training retrieval of memories of particular valences.

Our analysis also suggests that, compared to people without diagnoses, people with psychiatric diagnoses recall fewer specific memories and more general memories – that is, more memories for categories of events or extended events. These findings mean that we are unable to draw firm conclusions regarding whether these constructs are opposite sides of the same coin or whether they are separable constructs that correlate. However, it is of note that some studies have indicated that people with psychiatric diagnoses have difficulty retrieving both specific and general memories (Hitchcock, Rodrigues, et al., 2019; Piltan et al., 2020). These studies used an alternating instructions AMT (AMT-AI; Dritschel et al., 2014) wherein participants are asked to retrieve specific memories in some blocks of trials and to retrieve general memories in other blocks. The findings of these studies suggest that the abilities to retrieve specific and general memories may be separable and that some people may have problems in flexibly shifting between these levels of autobiographical memory when the situation demands it of them or when they are instructed to do so. Williams (2006) initially suggested that both specific and general memories are important as we may have to learn from specific autobiographical events (e.g., when I delivered a speech to my classmates) whilst also generalising the outcomes or solutions of these events to other related events in our present or future (e.g., to other public speaking events) or to our broader definition of who we are (e.g., I am confident public speaker). Our findings do not refute the idea that people with diagnoses have difficulty in flexibly shifting between specific and general memories so we would encourage further research in that area.

Contrary to our hypotheses, group differences in the CaRFAX process variables did not explain a significant amount of variance in effect sizes for group differences in specific or general memory retrieval. Although this finding appears to contrast with a previous narrative
review (Sumner et al., 2014) it accords with the findings of another review that focused on rAMS/OGM in child psychopathology and which also found little support for CaRFAx (Hitchcock et al., 2014). One finding that could, indirectly, support the contribution of CaRFAx processes to rAMS/OGM, regards our finding that participants with diagnoses had greater difficulty retrieving specific memories in tasks with shorter (30 second) retrieval times. This finding aligns with others that indicate that older adults without depression are significantly less likely to retrieve specific memories than younger non-depressed adults in paradigms that have shorter recall times (Barry et al., 2020). This is relevant because older adults have shown similar difficulties retrieving specific memories to depressed, younger people (Wilson & Gregory, 2018). It could be that people with diagnoses, as with older adults, lack the executive capacity to rapidly retrieve a specific memory within short periods of time but are able to retrieve such memories if they are given a longer time. However, it is unclear why the same (or inverse) finding was not present in our analysis of general memory retrieval. Future research could provide a more direct test of the contribution of retrieval time to between-group differences in specificity/generality.

There were limitations to our analysis, however, that prevent us from drawing firm conclusions about the contribution of CaRFAx variables to group differences in rAMS/OGM. First, most studies explored between group differences in specificity/generality and comparatively few additionally included measures of CaRFAx variables; the most analysed construct, rumination, was measured in only 14.3% of comparisons (k = 14; See Table 1 for a breakdown for each of specificity and overgenerality analyses). As such, these analyses were poorly powered. Second, of those that did measure these variables, there was little consistency in the exact measures that they employed, perhaps because of the broad nature of each CaRFAx construct. This is particularly problematic for the analysis of
avoidance and executive functioning as there are a number of ways that one can operationalise these constructs.

It must be repeated that our analysis of CaRFAX was not a test of the strength of the association between CaRFAX processes and rAMS/OGM broadly speaking but was instead an analysis of the contribution of group differences in these CaRFAX processes, to group differences in rAMS/OGM. If CaRFAX does contribute to rAMS/OGM, then normal variability in the extent of differences between clinical and control groups in these processes should be associated with differences in rAMS/OGM too. However, it is of note that some process variables might explain rAMS/OGM in clinical groups but not control groups (Farina et al., 2019) and our analysis strategy would hide such effects. Also, the CaRFAX process variables are expected to influence specificity/generality in interaction, not in isolation, but as most studies explored only one process variable, we were unable to explore these effects. An analysis of correlations between CaRFAX processes and rAMS/OGM was not possible within this analysis given the scarcity of analyses of CaRFAX processes within clinical-control analyses and, among the studies that did conduct such analyses there was limited reporting of within-group correlation coefficients.

As such, our analysis should not be used to discourage researchers from exploring the contribution of CaRFAX variables to rAMS/OGM but should motivate more studies to include measures of such variables and to fully report correlation coefficients within their clinical and control groups separately. This would increase the statistical power of future meta-analyses and enable researchers to directly test the association between CaRFAX and rAMS/OGM and to make clinical and control comparisons therein. Importantly, however, continued correlational analyses of CaRFAX variables will tell us little about causality. Rather, experimental manipulations of these variables (e.g., Sutherland & Bryant, 2007) or measuring specificity/generality in interventions for the other processes such as executive
functioning (e.g., Schweizer et al., 2011) are the gold standard that future studies must adhere to in order to move this field forward.

It is also possible that the reason we found poor support for the CaRFAX processes within rAMS/OGM is because of the limitations to the diagnoses used to separate the groups in each study. Conway et al. (2019) suggest that clinical-control comparisons, such as those sampled here, lose a considerable amount of information from people who do not fall neatly into the clinical or control group. This might explain why group differences in symptom severity did not contribute to effect size variance. Conway et al. (2019) recommend that future investigations should recruit participants on the basis of broader psychopathological dimensions shared by many diagnoses whilst also taking steps to ensure that one’s sample is representative of the spectrum of this dimension (e.g., inpatient and outpatient groups, community-dwelling treatment and non-treatment seeking groups etc.). They also call for the use of larger samples such that the relations between psychopathological constructs such as rAMS/OGM and particular symptom dimensions can be analysed with adequate power.

This approach was used, for example, in studies with participants from the Youth Emotion Project (YEP; Zinbarg et al., 2010) where participants across the range of the broad symptom dimension neuroticism were sampled, but where participants with higher levels of neuroticism were oversampled. This enabled researchers to conduct a comprehensive analysis of the contribution of the CaRFAX mechanisms to autobiographical memory specificity within a single investigation (Sumner et al., 2014). Sampling based on single diagnostic categories is only likely to hold this field back, and as such we recommend that inclusive sampling approaches such as those used by YEP should be utilised going forward. Such an analysis may yield an association between the severity of one’s psychopathology and their specificity/overgenerality, in accordance with other studies (Liu et al., 2013; Van Vreeswijk & De Wilde, 2004), but in contrast with the analysis presented here.
Nevertheless, it is possible that other moderators not examined within the studies sampled here might better contribute to the rAMS/OGM that is observed in clinical groups than the moderators analysed here. One notable omission from existing studies is a consideration of the social context in which memories are acquired, shared, and rehearsed. Future studies could explore the social factors that might contribute towards rAMS/OGM in psychiatric groups, such as social forms of reinforcement or punishment for sharing specific, detailed memories about ourselves (Debeer et al., 2014). In the studies sampled here, 82% of effects came from studies where participants had to report their memories verbally to an experimenter. If a person has fears of negative judgement, for example, due to a tendency to be self-critical or a history of being dismissed or criticised for sharing specific memories in the past, they may be reluctant to share such personal details to a stranger.

In addition, the CaRFAX model explains rAMS/OGM in terms of retrieval but says little about the way memories are encoded. The brain is constantly trying to use our previous experiences to construct an adaptive model of our world so that it can make predictions about, and prepare us for, our future (Van den Bergh et al., 2020). From this perspective rAMS/OGM is not just a retrieval phenomenon but it also directly influences, and is influenced by, the way that ongoing experiences are encoded. Experiments that manipulate cognition during event encoding, or longitudinal studies where a record of ongoing experiences is kept, perhaps with experience sampling methodology, could be used to examine the effects of processing and encoding on subsequent recall ability between people vulnerable to, or diagnosed with, clinical diagnoses, compared to diagnoses-free people.

The overview of methodological characteristics of the studies presented here enables us to make some other recommendations regarding the methodology of future investigations. The majority of studies used a standard AMT where participants were explicitly instructed to retrieve specific memories in response to five positive and five negative cue words. Most
studies also used the raw number of specific (67%) and general (82%) memories retrieved, as opposed to proportion scores. Also, 65% of studies that measured general memories included both categorical and extended memories within their definition and the remainder either used only categorical memories or did not specify what they meant by general memories. Each of these factors did not explain a significant amount of effect size variance, although the apparent scarcity of studies that used alternative cue-types or instruction sets may have limited our ability to analyse these moderating effects robustly.

In the future, different research questions will of course necessitate the use of different cuing paradigms and procedures and so we would not advocate for complete methodological consistency across the field. However, where possible, we urge researchers to use raw scores for the number of memories retrieved and when operationalising general memories, to report both categorical and extended memories or to combine the two. One might argue that as these factors did not contribute to variance in effect sizes it is not worth advocating for consistency in the way these methods are applied. However, if these methodological choices do not influence effect sizes, then it does not make sense for researchers to be inconsistent. In such cases, inconsistency only serves to confuse the naïve reader into thinking that different constructs are being investigated, or that slight changes are being made to methodology in order to p-hack. At the very least, studies should report exactly how they operationalise their outcome variables as, for example, a non-trivial number of studies used proportion scores but did not explicitly say what their scores were a proportion of (Specific: 16.5%; General: 5.4%).

Relatedly, there was a fairly high risk of bias across the studies sampled here, although the presence of suspected risk did not appear to be associated with between-group differences in the recall of specific or general memories. Nevertheless, future studies in this area should randomise both their presentation of their cuing task relative to other tasks within
their test protocol and the order of cues within their cuing task. They should also ensure that separate experimenters gather data and code memories and that both are blind to participants’ diagnostic status and the study hypotheses. Finally, in order to ensure open and transparent reporting, studies – their methodology, hypotheses, and data analysis strategies – should be pre-registered and data and analysis scripts should be made freely available during peer review and after publication.

Several variables were reported by some studies but not enough for us to provide a robust analysis of the contribution of these variables to variance in effect sizes. In particular, too few studies included data on illness duration, age of onset, medication use or ethnicity. Self-report data for each of these variables are easy enough to acquire from participants and would allow researchers to explore and confirm the representativeness of their sample to the broader clinical community. Among studies that did report ethnic information for their sample, participants were overwhelmingly white. Indeed, the vast majority of the studies sampled here come from America, Europe, and Australia. There is now a wealth of evidence attesting to how aspects of culture – such as those related to interdependence – can influence how autobiographical memories are shared and reported (Humphries & Jobson, 2012; Jobson, 2009; Wang et al., 2011, 2018; Wang & Conway, 2004). Although there is some evidence that the presence of psychiatric diagnoses such as PTSD can neutralise any cultural differences in the overall reporting of specific memories that might otherwise be expected (Jobson et al., 2014) there may still be cultural differences amongst people in the themes that their specific memories involve (e.g., the extent to which the memories are social in content; Jobson & Cheraghi, 2016).

**Limitations of the meta-analytic review**

Our meta-analytic review therefore possesses several limitations that bear highlighting. First, our analysis is heavily skewed towards comparisons with Major Depressive Disorder, with a
striking lack of studies amongst participants with Anxiety diagnoses. Second, our analysis of process moderators (e.g., rumination) is limited because of a relative lack of analyses and reporting of these variables within clinical-control comparisons and the substantial heterogeneity that exists in the way that these variables – particularly, executive functioning – have been operationalised. Demographic and clinical sample characteristics were also scarcely recorded or reported, limiting our ability to analyse these potentially important moderators. Relatedly, there was a substantial risk of bias within the available literature, mostly due to the lack of transparency with which methods, analytical procedures and results are reported and the lack of adherence to proper open science practices (e.g., pre-registration).

Although these limitations are concerned with the available literature itself, there are also limitations associated with our sampling of the literature. Our study was confined to English-language reports. It is of note that we did ask the listserv associated with this field for any additional data that they might have, and this was not constrained by any reporting language. No additional datasets were contributed. However, both this listserv and the published literature are notably Anglo-European-centric, with most non-English language researchers/studies being concentrated within European countries such as Spain (Ricarte et al., 2014), France (Bennouna-Greene et al., 2012) and Germany (Oertel-Knöchel et al., 2012) with few exceptions (Piltan et al., 2020). The nature of our search and inclusion criteria mean that it is unclear to what extent studies have been published in other languages that might have otherwise met our inclusion criteria. This means that the present meta-analytic review is biased towards Western-centric English language reporting.

Conclusion

Nevertheless, this multi-level meta-analytic review suggests that difficulty retrieving specific autobiographical memories and a tendency to instead retrieve general autographical
memories is a feature of a range of different psychiatric diagnoses. Our analysis of potential clinical, demographic, and methodological moderators did not offer any definitive answers as to what might explain the size of these group differences other than a suggestion that clinical participants have marked difficulty retrieving specific memories when the time given for retrieval is particularly short. We were also unable to confirm the contribution of clinical and control group differences in the seminal CaRFAx variables to group differences in rAMS/OGM. To take this field forward, we recommend that future research engages more diverse diagnostic groups so that the potential transdiagnosticity of rAMS/OGM can be further examined. We also recommend that future investigations in this area should engage in more consistent and transparent reporting of their data, and more diverse sampling of underrepresented groups and cultures, to reduce the risk of bias that is evident in this field. Finally, we hope that our analysis is a starting point and that future investigators take advantage of the data that has been collated here and which is now available to them to facilitate future meta-analyses in this area.
Table 1. Study characteristics

<table>
<thead>
<tr>
<th>Diagnoses</th>
<th>Specificity (k = 85)</th>
<th>General (k = 56)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k</td>
<td>%</td>
</tr>
<tr>
<td>Mood Disorders</td>
<td>37</td>
<td>43.5</td>
</tr>
<tr>
<td>Psychosis</td>
<td>11</td>
<td>12.9</td>
</tr>
<tr>
<td>Trauma-related</td>
<td>11</td>
<td>12.9</td>
</tr>
<tr>
<td>Borderline</td>
<td>8</td>
<td>9.4</td>
</tr>
<tr>
<td>Anxiety</td>
<td>6</td>
<td>7.1</td>
</tr>
<tr>
<td>Other</td>
<td>12</td>
<td>14.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement</th>
<th>k</th>
<th>%</th>
<th>k</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMT</td>
<td>64</td>
<td>75.3</td>
<td>46</td>
<td>82.1</td>
</tr>
<tr>
<td>Verbal</td>
<td>68</td>
<td>80.0</td>
<td>49</td>
<td>87.5</td>
</tr>
<tr>
<td>Raw number</td>
<td>54</td>
<td>63.5</td>
<td>29</td>
<td>51.8</td>
</tr>
<tr>
<td>Specific Instructions</td>
<td>67</td>
<td>78.8</td>
<td>48</td>
<td>85.7</td>
</tr>
<tr>
<td>60 Seconds</td>
<td>54</td>
<td>63.5</td>
<td>29</td>
<td>51.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Process variables</th>
<th>k</th>
<th>%</th>
<th>k</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rumination</td>
<td>14</td>
<td>16.5</td>
<td>11</td>
<td>19.6</td>
</tr>
<tr>
<td>Avoidance</td>
<td>9</td>
<td>10.6</td>
<td>8</td>
<td>14.3</td>
</tr>
<tr>
<td>Executive Functioning</td>
<td>11</td>
<td>12.9</td>
<td>8</td>
<td>14.3</td>
</tr>
<tr>
<td>Fluency</td>
<td>9</td>
<td>10.6</td>
<td>6</td>
<td>10.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinical characteristics</th>
<th>k</th>
<th>M (SD)</th>
<th>k</th>
<th>M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>80</td>
<td>36.89 (11.81)</td>
<td>52</td>
<td>33.28 (13.82)</td>
</tr>
<tr>
<td>Female</td>
<td>78</td>
<td>0.64 (0.23)</td>
<td>49</td>
<td>0.64 (0.22)</td>
</tr>
<tr>
<td>White</td>
<td>10</td>
<td>0.71 (0.25)</td>
<td>8</td>
<td>0.82 (0.18)</td>
</tr>
<tr>
<td>Years educated</td>
<td>31</td>
<td>10.19 (3.36)</td>
<td>20</td>
<td>9.93 (3.54)</td>
</tr>
<tr>
<td>Medicated</td>
<td>18</td>
<td>0.80 (0.28)</td>
<td>8</td>
<td>0.75 (0.39)</td>
</tr>
<tr>
<td>Illness duration</td>
<td>11</td>
<td>162.00 (89.58)</td>
<td>4</td>
<td>100.23 (104.23)</td>
</tr>
<tr>
<td>Illness onset</td>
<td>6</td>
<td>21.87 (1.10)</td>
<td>3</td>
<td>20.66 (0.63)</td>
</tr>
</tbody>
</table>

Note. Table of the characteristics of the sampled studies for each of the analyses of memory specificity and overgenerality. The breakdown of the number of effects (k; and the percent of the total number of effects for each analysis that this represents) for each diagnostic category is given. In addition, we present several key measurement characteristics, including the number of effects that used a traditional Autobiographical Memory Test (AMT), the number required that participants to respond verbally (Verbal), the number that used a raw total score for specificity/generality (Raw number), the number that specified to participants that they should recall a specific memory (Specific Instructions) and the number that gave participants 60 seconds to retrieve and report their memory (60 seconds). We also give the number of effects that included a measurement of each process variable. Finally, means (M) and standard deviations (SD) are given for the characteristics of the clinical samples across the sampled studies, including their mean age, the proportion of females, the proportion who identified as white, the number of years educated, the proportion who were taking medication, the mean illness duration in months and the mean age (in years) of illness onset.
Table 2. Pooled effect sizes by diagnostic group

<table>
<thead>
<tr>
<th></th>
<th>Specific memories</th>
<th></th>
<th></th>
<th></th>
<th>Overgeneral memories</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>k</td>
<td>g</td>
<td>SE</td>
<td>lbound</td>
<td>ubound</td>
<td>Z</td>
<td>P</td>
<td>n</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>12</td>
<td>-1.281</td>
<td>0.156</td>
<td>-1.587</td>
<td>-0.975</td>
<td>-8.209</td>
<td>0.001</td>
<td>10</td>
</tr>
<tr>
<td>Anxiety</td>
<td>5</td>
<td>6</td>
<td>-0.443</td>
<td>0.321</td>
<td>-1.073</td>
<td>0.187</td>
<td>-1.379</td>
<td>0.168</td>
<td>2</td>
</tr>
<tr>
<td>Mood</td>
<td>33</td>
<td>37</td>
<td>-0.659</td>
<td>0.087</td>
<td>-0.829</td>
<td>-0.490</td>
<td>-7.619</td>
<td>0.001</td>
<td>24</td>
</tr>
<tr>
<td>BPD</td>
<td>5</td>
<td>8</td>
<td>-0.587</td>
<td>0.132</td>
<td>-0.847</td>
<td>-0.328</td>
<td>-4.442</td>
<td>0.001</td>
<td>3</td>
</tr>
<tr>
<td>PTS</td>
<td>11</td>
<td>11</td>
<td>-1.193</td>
<td>0.435</td>
<td>-2.045</td>
<td>-0.341</td>
<td>-2.743</td>
<td>0.006</td>
<td>6</td>
</tr>
<tr>
<td>Psychosis</td>
<td>11</td>
<td>10</td>
<td>-0.982</td>
<td>0.211</td>
<td>-1.395</td>
<td>-0.568</td>
<td>-4.648</td>
<td>0.001</td>
<td>4</td>
</tr>
<tr>
<td>All</td>
<td>64</td>
<td>85</td>
<td>-0.864</td>
<td>0.085</td>
<td>-1.030</td>
<td>-0.698</td>
<td>-10.223</td>
<td>0.001</td>
<td>44</td>
</tr>
</tbody>
</table>

Note. Pooled effect sizes (g), standard error (SE), low and upper 95% intervals (lbound; ubound), Z scores and P values for each diagnostic group (and across all groups) for differences between people with and without psychiatric diagnoses in the retrieval of specific (left) and overgeneral (right) memories. N = number of studies; k = number of effects; g = Hedges’ g; BPD = Borderline Personality Disorder; PTS = Posttraumatic Stress.
Figure 1. Flow chart for study inclusion

Note. Flow diagram of study identification, screening, eligibility, and inclusion. Information on the number of studies (n) and that were identified at each step of the search and screening process, descending downwards to the final number of included studies, after exclusion, and their associated effect sizes (k).
Note. A forest plot of effect sizes and 95% confidence intervals (CI) for studies that measured differences between people with and without psychiatric diagnoses in the retrieval of specific autobiographical memories. Participants primary diagnosis and the sample size (N) are also given. The lower most diamond represents the pooled effect size across studies. Measures of heterogeneity (Q, and its associated significance test, and I²) are also provided.

N = number of participants; ASD = Acute Stress Disorder; PTSD = Posttraumatic Stress Disorder; SMD = Standardized Mean Difference (Hedge’s g).
Figure 3. Forest plot – general memories

Note. A forest plot of effect sizes and 95% confidence intervals (CI) for studies that measured differences between people with and without psychiatric diagnoses in the retrieval of general autobiographical memories. Participants primary diagnosis and the sample size (N) are also given. The lower most diamond represents the pooled effect size across studies.

Measures of heterogeneity (Q, and its associated significance test, and I²) are also provided.

ASD = Acute Stress Disorder; PTSD = Posttraumatic Stress Disorder; SMD = Standardized Mean Difference (Hedge’s g).
Figure 4. Funnel plots

Note. Funnel plots for relationship between sample variance and effect sizes (Hedges’ $g$; Left: Specific memories; Right: General memories) with contours for different significance levels.
Figure 5. Risk of bias

Note. Traffic light plot of the percentage of studies that were marked as having a low risk of bias, some concern, or a high risk of bias for each of the six categories of bias.
References

*Studies included within the meta-analysis


https://doi.org/10.1016/j.brat.2019.02.001

https://doi.org/10.1080/02699931.2013.807776

https://doi.org/10.1016/j.concog.2011.10.006

https://doi.org/10.1521/pedi_2018_32_368

https://doi.org/10.1093/schbul/sbv099


https://doi.org/10.1037/0033-295X.107.2.261

https://doi.org/10.1017/S0033291703007529


https://doi.org/10.1023/A:1005527703011


Hallford, D. J., Noory, N., & Mellor, D. (2018). Reduced autobiographical memory specificity as a mediating factor between general anxiety symptoms and performance on


National Comorbidity Survey Replication. *Arch Gen Psychiatry*, 62, 593–602.


mental disorders: A comprehensive analysis based on 145 990 survey respondents from 27 countries. *Epidemiology and Psychiatric Sciences.*

https://doi.org/10.1017/S2045796020000633


https://doi.org/10.1016/j.psychres.2009.10.004


https://doi.org/10.1016/j.brat.2012.03.009


https://doi.org/10.1002/jclp.22746


https://doi.org/10.1017/bec.2013.17


*Pollock, L. R., & Williams, J. M. G. (2001). Effective Problem Solving in Suicide Attempters Depends on Specific Autobiographical Recall. *Suicide and Life-Threatening Behavior.* https://doi.org/10.1521/suli.31.4.386.22041


https://doi.org/10.1016/j.schres.2014.10.027


Williams, J. M. G., Barnhofer, T., Crane, C., Hermans, D., Raes, F., Watkins, E., &


Young, K. D., Erickson, K., & Drevets, W. C. (2012). Match between Cue and Memory Valence during Autobiographical Memory Recall in Depression. *Psychological Reports,*
