Learning new word meanings from story reading: The benefit of immediate testing

Rachael C Hulme [Corresp., 1, 2] , **Jennifer M Rodd** [1]

[1] Department of Experimental Psychology, University College London, University of London, London, United Kingdom

[2] Aston Institute of Health and Neurodevelopment and School of Psychology, Aston University, Birmingham, United Kingdom

Corresponding Author: Rachael C Hulme
Email address: rachael.hulme.14@ucl.ac.uk

This study investigated how word meanings can be learned from natural story reading. Three experiments with adult participants compared naturalistic incidental learning with intentional learning of new meanings for familiar words, and examined the role of immediate tests in maintaining memory of new word meanings. In Experiment 1, participants learned new meanings for familiar words through incidental (story reading) and intentional (definition training task) conditions. Memory was tested with cued recall of meanings and multiple-choice meaning-to-word matching immediately and 24 hours later. Results for both measures showed higher accuracy for intentional learning, which was also more time efficient than incidental learning. However, there was reasonably good learning from both methods, and items learned incidentally through stories appeared less susceptible to forgetting over 24 hours. It was possible that retrieval practice at the immediate test may have aided learning and improved memory of new word meanings 24 hours later, especially for the incidental story reading condition. Two preregistered experiments then examined the role of immediate testing in long-term retention of new meanings for familiar words. There was a strong testing effect for word meanings learned through intentional and incidental conditions (Experiment 2), which was non-significantly larger for items learned incidentally through stories. Both cued recall and multiple-choice tests were each individually sufficient to enhance retention compared to having no immediate test (Experiment 3), with a larger learning boost from multiple-choice. This research emphasises (i) the resilience of word meanings learned incidentally through stories and (ii) the key role that testing can play in boosting vocabulary learning from story reading.

4 # Learning new word meanings from story reading:

5 # The benefit of immediate testing

9 Rachael C. Hulme[1,2] and Jennifer M. Rodd[1]

10 [1]Department of Experimental Psychology, University College London, London, UK

11 [2]Aston Institute of Health and Neurodevelopment and School of Psychology, Aston University,

12 Birmingham, UK

14 Corresponding author:

15 Rachael Hulme[1,2]

16 Email address: rachael.hulme.14@ucl.ac.uk

17   **Abstract**

18   This study investigated how word meanings can be learned from natural story reading.

19   Three experiments with adult participants compared naturalistic incidental learning with

20   intentional learning of new meanings for familiar words, and examined the role of immediate

21   tests in maintaining memory of new word meanings. In Experiment 1, participants learned new

22   meanings for familiar words through incidental (story reading) and intentional (definition

23   training task) conditions. Memory was tested with cued recall of meanings and multiple-choice

24   meaning-to-word matching immediately and 24 hours later. Results for both measures showed

25   higher accuracy for intentional learning, which was also more time efficient than incidental

26   learning. However, there was reasonably good learning from both methods, and items learned

27   incidentally through stories appeared less susceptible to forgetting over 24 hours. It was possible

28   that retrieval practice at the immediate test may have aided learning and improved memory of

29   new word meanings 24 hours later, especially for the incidental story reading condition. Two

30   preregistered experiments then examined the role of immediate testing in long-term retention of

31   new meanings for familiar words. There was a strong testing effect for word meanings learned

32   through intentional and incidental conditions (Experiment 2), which was non-significantly larger

33   for items learned incidentally through stories. Both cued recall and multiple-choice tests were

34   each individually sufficient to enhance retention compared to having no immediate test

35   (Experiment 3), with a larger learning boost from multiple-choice. This research emphasises (i)

36   the resilience of word meanings learned incidentally through stories and (ii) the key role that

37   testing can play in boosting vocabulary learning from story reading.

39   *Keywords*

40   incidental learning; intentional learning; testing effect; homonyms; story reading

Introduction

42   Recent research emphasises the importance of good vocabulary knowledge: individuals

43   with better vocabulary perform better on reading comprehension tests, and have better

44   educational outcomes (Armstrong et al., 2017; Cain & Oakhill, 2014). Despite widespread

45   acceptance that incidental learning from natural linguistic environments (e.g., conversations,

46   books, TV) is the main source of vocabulary learning (Batterink & Neville, 2011; Nagy et al.,

47   1985, 1987), most studies of vocabulary learning in adults use highly artificial stimuli, tasks, and

48   learning conditions. The current experiments focus on learning from naturalistic fiction stories

49   that are read by participants in their native language without any explicit instructions to learn the

50   new vocabulary that the stories contain. We investigate the extent to which people's ability to

51   retain newly-learned word meanings over time is improved by requiring them to retrieve these

52   word meanings during the intervening period between encoding and a later test. The presence of

53   such a beneficial "testing effect" has been well established through studies of explicit, intentional

54   learning (for reviews, see Roediger & Butler, 2011; Rowland, 2014), but it is unclear whether

55   retrieval would similarly enhance memory for vocabulary learned under more naturalistic,

56   incidental learning conditions. The finding that vocabulary learning from naturalistic materials

57   could be significantly boosted by a brief episode of testing could provide a simple approach to

58   boosting vocabulary gains in real-world settings.

59   Incidental vocabulary learning is defined as learning words and their meanings whilst

60   engaged in another activity such as listening or reading for comprehension (Hulstijn, 2003). A

61   real-life context in which adults often learn new words and their meanings is when reading

62   fiction, due to the rich and varied situations that are often depicted (Nation, 2017). Studies of

63   word learning from stories by adult native-language (L1) readers have adopted highly naturalistic

64   methods by using either authentic texts (Godfroid et al., 2017; Saragi et al., 1978) or texts

65   modified or written specifically for the purposes of the studies (Batterink & Neville, 2011;

66   Henderson et al., 2015; Pellicer-Sánchez, 2016). In these studies participants read works of

67   fiction with the primary focus being on comprehension, with vocabulary learning as a by-

68   product. To discourage intentional learning strategies, readers are not given any instruction to

69   learn new vocabulary encountered in a text and are not informed that their memory will later be

70   tested.

71   The current study uses a paradigm developed by Hulme et al. (2019) in which

72   participants encounter artificial new meanings for familiar English words in the context of

73   custom-written short stories (e.g., learning that a *foam* is a type of safe concealed within a piece

74   of furniture). This ability to learn new word meanings is a key aspect of vocabulary

75   development: around 80% of common English words have more than one definition (Rodd,

76   2018; Rodd et al., 2002). Adults often learn additional word senses/meanings, and continue to

77   update their knowledge of these words throughout their adult lives (Betts et al., 2018; Gaskell et

78   al., 2019; Gilbert et al., 2018, 2021; Rodd et al., 2013). Examples of reasons why adults learn

79   new meanings for familiar words include language evolution (e.g., the internet-related meaning

80   of "troll"), or learning a new subject or activity (e.g., the sailing term "boom"; Eligio & Kaschak,

81   2021; Rodd et al., 2012, 2016). New meanings are often learned when reading stories, especially

82   of the science fiction or fantasy genres (e.g., a "galleon" is a coin of the wizarding currency in

83   the Harry Potter series of novels by J. K. Rowling). Recently, Fang et al. (2016) proposed that

84   learning new meanings for familiar words is a dual-phase process whereby familiarity with the

85   word form may facilitate learning with the initial encounters, but inhibition due to meaning

86   competition begins to take effect after subsequent exposures to the newly ambiguous word

87   (Maciejewski et al., 2020; Maciejewski & Klepousniotou, 2020; see Rodd, 2020 for review).

88   Hulme et al. (2019) found that participants were able to recall the new meanings for the

89   known words reasonably well (38.5% correct) after only two exposures in a story context, with a

90   linear increase in meaning recall with additional exposures (63.5% correct after eight exposures).

91   Interestingly, Hulme et al.'s (2019) participants showed no significant forgetting of the new

92   meanings they had learned at a surprise test one week later across all of the exposure conditions.

93   The current study further examines this incidental form of word learning by (i) comparing

94   performance to a more explicit learning condition and (ii) investigating the potential boost to

95   performance from an immediate test of knowledge after training. Understanding how these two

96   factors impact on long-term retention of vocabulary will provide a critical foundation for

97   subsequent development of interventions to boost vocabulary acquisition.

98   The conditions of initial vocabulary acquisition (incidental or intentional) prompt

99   different types of information processing, which may affect retention of word meanings in

100 different ways. Vocabulary learned under intentional conditions may be retained better over time

101 because more attention is directly focussed on encoding the word meanings, and the meaning is

102 made more explicit. This more strategic processing might be particularly important for

103 facilitating access to prior knowledge in the case of learning new ambiguous words where the

104 learner may benefit from more explicitly noticing the mismatch between the familiar word

105 meaning and the new meaning. In contrast, incidental vocabulary learning from story reading

106 may benefit from the rich and informative story contexts (Webb, 2008), and it has been

107 suggested that the increased mental effort required to encode new word meanings inferred from

108 context may be beneficial for retention (Hulstijn, 1992). However, it is also important to consider

109   that while incidental vocabulary learning is usually contextualised (with words embedded in

110   informative contexts from which meaning is inferred), intentional vocabulary learning may also

111   involve context, or it can be decontextualised.

112   The consensus from the literature on adult second language (L2) learning (e.g., Hulstijn,

113   1992; Lehmann, 2007; Peters et al., 2009), and research with teenagers learning L1 vocabulary

114   (Konopak et al., 1987) is that intentional learning offers greater vocabulary gains and is more

115   efficient than incidental learning. However, some other studies have found little difference

116   (Lehmann, 2007), or even an efficiency advantage in terms of words learned per minute for

117   incidental learning (Mason & Krashen, 2004). Several recent studies with adult L1 readers have

118   also found good levels of native language vocabulary acquisition from reading alone (Batterink

119   & Neville, 2011; Godfroid et al., 2017; Pellicer-Sánchez, 2016). A further key factor that could

120   differ between vocabulary acquisition under incidental and intentional learning conditions is the

121   impact of testing on subsequent retention.

122   The "testing effect" refers to the finding that testing memory following training can

123   enhance long-term retention, as the additional retrieval practice at test affords an opportunity for

124   further learning (for reviews, see Roediger & Butler, 2011; Rowland, 2014). The effect has been

125   demonstrated as robust in various experiments using explicit, intentional learning conditions.

126   However, it is unclear whether the testing effect would provide a similar benefit for vocabulary

127   learned under incidental learning conditions. Given that the vast majority of native language

128   words and their meanings are learned incidentally (Batterink & Neville, 2011), it is important to

129   examine the impact of the testing effect under such learning conditions. If the presence of a

130   quick, immediate vocabulary test can indeed enhance learning/retention for incidentally learned

131   vocabulary this could potentially provide a simple method for boosting vocabulary gains from

132   story reading, especially within educational settings.

133   In vocabulary learning research, retrieval practice has been shown to lead to better

134   retention of new words over time with adults learning second language (L2) vocabulary under

135   intentional conditions (e.g., Fritz et al., 2007; Karpicke & Roediger, 2008; van den Broek et al.,

136   2013, 2018), and similarly with children learning novel L1 words (Goossens et al., 2014;

137   Goossens et al., 2014; Toppino & Cohen, 2009). The testing effect further enhances retention

138   when feedback is provided on performance on the immediate test (e.g., Pashler et al., 2005), but

139   retrieval practice is often beneficial even in the absence of any feedback (Roediger & Butler,

140   2011). The precise neurocognitive mechanism underlying the testing effect is currently unclear,

141   but it has recently been suggested that retrieval practice may provide a fast track to consolidation

142   of new information through the online reactivation of related knowledge (Antony et al., 2017;

143   see the General Discussion for further discussion).

144   It is possible that different learning conditions preceding retrieval practice could

145   moderate the testing effect for various reasons. For example, it is thought that semantic

146   elaboration may be key to the neurocognitive mechanism underlying the testing effect

147   (Carpenter, 2009). If this is the case, then the richer story contexts during encoding in the

148   incidental condition could provide more fertile material for semantic elaboration, thus enhancing

149   the testing effect. On the other hand, research has suggested that the benefits of retrieval practice

150   are greater when retrieval success during practice is high (Rowland, 2014). Therefore, if

151   intentional learning is more effective than incidental learning then this could lead to a stronger

152   testing effect following encoding under intentional learning conditions.

153   Retrieval practice has been shown to benefit long-term retention of information learned

154   under a variety of conditions (e.g., Butler, 2010; Karpicke & Roediger, 2008; Roediger &

155   Karpicke, 2006a; van den Broek et al., 2013), although little research has compared across

156   different learning conditions. One study (Goossens et al., 2014) directly compared the impact of

157   testing on children's learning of novel L1 vocabulary from a story context to learning new words

158   in isolation. Results showed that children correctly recalled more word meanings that had been

159   tested, and children in the word list condition remembered the word meanings better overall than

160   those in the story condition. The testing effect was also slightly stronger for the word list

161   condition. However, learning was not incidental in either condition in this study, and children

162   who heard the story also had the meanings of the words explained to them. Furthermore, the

163   participants in this study were children (aged 8-11), and results may differ for adults whose

164   advanced language skills and vocabulary knowledge make them better equipped to learn more

165   successfully from the richer contexts that stories provide. It therefore remains to be seen whether

166   the benefit of retrieval practice would differ for the learning of new word meanings acquired

167   solely under incidental conditions in a story context, as compared with learning under intentional

168   conditions.

169

170   **Experiment 1: Incidental versus intentional learning**

171   Experiment 1 compared the story-reading method designed by Hulme et al. (2019) for

172   studying incidental learning of new meanings for familiar words with a more conventional,

173   intentional training procedure. This provided a baseline assessment of how well adults are able to

174   learn new word meanings from a naturalistic incidental learning paradigm as compared to a more

175   conventional explicit approach to vocabulary learning, and provided a foundation for the

176   subsequent preregistered experiments to investigate learning performance in more detail.

177   Specifically, Experiments 2 and 3 will follow up on Experiment 1 to examine whether the

178   inclusion of an immediate test of new vocabulary knowledge aids learning and improves

179   memory of new word meanings 24 hours later. This may be especially pertinent for vocabulary

180   acquired through incidental learning conditions as it may prompt participants to adopt different

181   information processing strategies after initial acquisition.

182   In Experiment 1 participants learned novel meanings for existing unambiguous words

183   through both incidental story-reading (as in the study by Hulme et al., 2019) and a newly

184   developed intentional task-based learning procedure, with the same number of exposures to

185   items. The two learning conditions were implemented based on typical paradigms for these two

186   types of learning. However, it is important to note that there are multiple differences in the

187   learning experience, for example only the incidental learning paradigm required participants to

188   infer meaning from context. While it is more common for incidental learning to be

189   contextualised in this way, some intentional learning paradigms also involve contextualised

190   learning (see for example: van den Broek et al., 2018). The stories used in the incidental learning

191   condition combined naturalistic elements of authentic texts (Godfroid et al., 2017; Saragi et al.,

192   1978) with precise experimental control over the exposure to items within the text (Batterink &

193   Neville, 2011; Pellicer-Sánchez, 2016). Items were encountered incidentally within the stories

194   that participants read for comprehension and were central to the narrative.

195   Participants' knowledge of the new meanings for all items was assessed first through

196   cued recall, and second through a multiple-choice meaning-to-word matching test. The recall

197   measure is a harder test with fewer cues to help retrieve memories of the new word meanings,

198   while the multiple-choice test is a recognition measure with more cues and is therefore the easier

199 of the two tests. Using two tests of learning with different difficulty levels allowed us to reduce

200 the possibility of floor/ceiling effects. The tests were administered both immediately after

201 learning, and again 24 hours later to assess longer-term retention. Based on the previous

202 research, we predicted that learning of new word meanings would be better for intentional

203 learning conditions, although we expected reasonably good vocabulary learning for the

204 incidental learning conditions in line with the findings of Hulme et al. (2019). We also predicted

205 there would be little forgetting after 24 hours, although we had no specific predictions as to

206 whether this would differ for new word meanings acquired through incidental or intentional

207 learning conditions. Our predictions were the same for the cued recall test the and multiple-

208 choice meaning-to-word matching test.

209   The materials, data, and analysis scripts for Experiment 1 can be found on the Open

210 Science Framework (OSF; https://osf.io/k32tw). For all experiments we report all measures,

211 conditions, data exclusions, and how we established the sample size.

212

213 **Method**

214   **Participants.**

215   We aimed to recruit 40 participants for Experiment 1. The study by Hulme et al. (2019)

216 included 64 participants who were trained on four items (one per exposure condition) in one of

217 16 experiment versions (four participants per version). In this study participants were trained on

218 eight items (four items per learning condition) in one of eight experiment versions (five

219 participants per version), we therefore expected power to be comparable to that of Hulme et al.

220 (2019). Forty participants were included in the experiment (age: $M = 30.1$ years, $SD = 7.1$; 23

221 female). Participants were recruited through the Prolific recruitment website (Damer & Bradley,

222   2014) using pre-screening criteria. They gave their informed consent before taking part (by

223   means of ticking boxes in the online consent form). The UCL Experimental Psychology Ethics

224   Committee granted ethical approval for the research (Ref: EP/2017/009). Participants were

225   invited to take part if they were a current UK resident, a monolingual native speaker of British

226   English, and had no diagnosis of reading or language impairments. They were paid for their

227   participation in the first session of the experiment (£5) and additionally upon completion of the

228   second session 24 hours later (£1). Of the 40 participants who completed the first session, 31 also

229   completed the 24-hour follow-up session on time (77.5%). One additional participant was

230   excluded from the second session due to completing it after the deadline (within six hours of

231   receiving the invitation for the follow-up session).

232   Five additional participants were excluded from the study—two were not monolingual

233   native British English speakers, and three got more than one of the multiple-choice

234   comprehension questions wrong when reading the stories (see Procedure). Excluded participants

235   were replaced during recruitment.

236

237   **Materials.**

238   ***Novel word meanings.***

239   The stimuli were 16 real English nouns that were given artificial new meanings, taken

240   from the study by Hulme et al. (2019) (see Table S1 for the stimuli: https://osf.io/m4wxa). The

241   new meanings were unrelated to the existing meanings of the words, and described hypothetical

242   innovations, discoveries, and inventions. There was one definition sentence for each of the

243   stimulus words that described its new meaning, for example: "A foam is a safe that is

244   incorporated into a piece of furniture with a wooden panel concealing the key lock, and each is

245   individually handcrafted so that no intruders are able to recognize the chief use of the furniture."

246   The sentences were matched for length ($M = 32.9$ words, $SD = 3.7$). Each new meaning had three

247   distinguishing semantic features to maintain a similar level of complexity for each new concept,

248   for example, for *foam*: "a safe inside a piece of furniture," "has a hidden key lock," and

249   "individually handcrafted to fool intruders." The words and their meanings were incorporated

250   into story narratives for the incidental learning condition, and the definition sentences were

251   presented to participants in the definition reading phase of the intentional learning condition.

252   Three shorter paraphrased excerpts of the definition sentences were created for use in the

253   two-alternative multiple-choice training task (length: $M = 11.29$ words; $SD = 2.13$), with each

254   sentence describing a semantic feature of that item (e.g., for "foam": "A secure place to store

255   valuables within an item of furniture."; "A safe with a wooden panel disguising the key lock.";

256   and "A bespoke handcrafted piece of furniture containing a safe hidden from intruders.").

257   Paraphrased versions were used to encourage participants to read the whole sentence each time,

258   rather than relying on recognition of the first words (see Table S2 for the short sentences:

259   https://osf.io/m4wxa).

260   An additional longer paraphrased version of each of the definition sentences (which were

261   used in the test of cued recall of word forms in Hulme et al.'s (2019) study) were used in the

262   multiple-choice test at the end of this experiment (see Table S3 for the sentences used for the

263   multiple-choice test: https://osf.io/m4wxa).

264

265   ***Short stories.***

266   The four short stories from Hulme et al.'s (2019) study were used to present stimuli to

267   participants in the incidental learning condition in this experiment (see the Supplementary

268   Materials for the stories: https://osf.io/m4wxa). These stories (ranging 2307-2446 words in

269   length) were written by a professional children's author (Story 1: *Pink Candy Dream*), and an

270   unpublished author (Story 2: *Prisons*, Story 3: *Reflections upon a Tribe*, and Story 4: *The Island*

271   *and Elsewhere*), and were designed to be interesting for an adult audience. Each story

272   incorporated four of the items in the context of their new meanings, with each item appearing a

273   total of eight times at naturally distributed positions within a story. No item appeared in more

274   than four consecutive sentences, and all items occurred on at least two different pages of the

275   story. On the first presentation of a stimulus word, sufficient information was given to allow the

276   reader to derive the new meaning from the context from the first exposure, for example, "'Yes,' I

277   murmured, breathing again. 'I *knew* it! It's a foam.' The ornate *chaise longue* was no ordinary

278   piece of furniture but concealed a built-in safe with an intricate key-operated locking system."

279   The amount of information about each new meaning in subsequent exposures varied naturally

280   with the story narratives. A degree of inference was required to extract the meaning from the

281   context to reflect natural word learning from reading where explicit definitions are rarely given.

282

283   **Design.**

284   The experiment employed a within-participants and within-items design: participants

285   were trained on four items through the incidental learning condition and four items through the

286   intentional learning condition. Each participant was trained on only half the total number of

287   stimuli as this was deemed to be a feasible number of new meanings to learn in a single session.

288   To ensure each new word meaning was seen an even number of times in each condition, and that

289   the order of learning conditions was counterbalanced across participants (to minimise any order

290   effects), we created eight versions of the experiment. Participants were pseudorandomly assigned

291   to one of the eight versions of the experiment. Time of test (immediate versus 24 hours later) was

292   also within-participants (based on the 31 participants who completed both sessions). The

293   dependent measures were accuracy in cued recall of the new word meanings, and accuracy in the

294   multiple-choice test.

295

296   **Procedure.**

297   The experiment was run online using Qualtrics (Qualtrics, 2015), and participants were

298   instructed to complete each session in one sitting without breaks. Participants were asked to read

299   the stories and definitions carefully, and were not told that their memory for the new word

300   meanings would be tested. Participants were told that the aim of the experiment was to study

301   subjective reading style and comprehension. After completing the first session, participants were

302   also not informed that they would be contacted at the same time the following day to invite them

303   to complete the second session to discourage the use of deliberate memorisation strategies.

304   Participants read one of the short stories in the incidental learning condition. Each story

305   was divided into five pages of roughly even length and displayed on-screen one page at a time.

306   After each page, a multiple-choice comprehension question appeared on a separate screen asking

307   about details of the story's plot from the preceding page (without probing details of the novel

308   word meanings). Participants were instructed to read the story closely, and to answer a multiple-

309   choice comprehension question after each page. Participants were able to re-read sentences on

310   the current page, although no instructions were given to participants on this; they were not able

311   to go back to reread previous pages. For the comprehension questions participants had to select

312   the correct answer from four options that appeared in a randomised order; they were designed to

313   be very easy for any participant who had attentively read the text. Participants were excluded if

314   they got more than one of the comprehension questions wrong.

315   The intentional learning condition consisted of two phases which both repeated once:

316   definition sentence reading, followed by two-alternative multiple-choice meaning-to-word

317   matching. In the definition reading phase, participants were presented with sentences that

318   described the key semantic features of each of the novel word meanings, stating the word to

319   which it referred. These four definition sentences were presented one at a time on separate pages,

320   and the order of presentation was randomised for each participant. Participants were instructed to

321   read each definition carefully to make sure they understood it before proceeding to the next one.

322   Once participants had read all of the definition sentences once, they moved immediately

323   on to the two-alternative multiple-choice meaning-to-word matching task. Participants were

324   presented one at a time with three shortened, paraphrased versions of the definitions of each of

325   the novel meanings. For each item participants were instructed to choose the correct word for the

326   new meaning from two options: the correct word and one foil word. After selecting one of the

327   options, participants were provided with feedback, which either said "Correct answer!" or

328   "Incorrect.". The short sentences were presented in a pseudorandomised order, ensuring that the

329   sentences referring to each item were roughly evenly spaced, and none referring to the same item

330   occurred one after another. The foil word for each trial was one of the other words from the

331   intentional training condition. Each foil word was paired an even number of times with each

332   correct word, and the order that the correct word and foil appeared in was randomised for each

333   trial. The two phases of the intentional training were then repeated in the same order. This gave a

334   total of two exposures to the novel word meanings from the definition sentence reading phase

335   and six exposures to the new meanings from the multiple-choice task, totalling eight exposures—

336   equal to the number of exposures in the incidental learning condition. Participants spent more

337   time reading the story (including comprehension questions; M = 12 mins 30 secs, SD = 4 mins

338   34 secs) than they spent on the intentional training task (M = 5 mins 28 secs, SD = 2 mins 55

339   secs), $t(39)$ = 11.43, p < .001.

340   After they had completed training through both the incidental and intentional learning

341   conditions, participants completed a brief filler task. This was the 34-item version of the Mill

342   Hill vocabulary test (Mill Hill Vocabulary Test, Set A: Multiple Choice: Buckner et al., 1996;

343   Raven, Raven, & Court, 1998). For each item, participants were required to select one word

344   from a list of six options that most closely matched the meaning of the presented word. None of

345   the stimulus items appeared in the filler task. The purpose of this task was to counteract any

346   recency effects of memory for stimulus items encountered toward the end of training; responses

347   were not analysed.

348   Participants were next given a cued recall test of all eight of the new meanings they had

349   encountered in the experiment when presented with the words. Participants saw each of the eight

350   words they had been trained on and were asked to recall the appropriate new meaning and type it

351   into a text box. They were encouraged to provide as much detail as possible and to try to answer

352   in full sentences even if they were unsure of their answer. If they could not remember anything

353   about the new meaning for the word, they were instructed to type "don't know." The order of

354   presentation of the words was randomised for each participant, with the four words from each

355   training method randomly intermixed within the test; this was also the case for the subsequent

356   test.

357   The second test was an eight-alternative multiple-choice meaning-to-word matching test.

358   Participants were presented one at a time with paraphrased definitions of the novel word

359   meanings they had been trained on. The sentences omitted the words to which they were

360   referring, and for each novel meaning participants were asked to select the word that they

361   thought matched the definition from a list of all eight of the stimulus words they had encountered

362   throughout the experiment. The order of the eight words to choose from was randomised for each

363   test item, and the order of presentation of the new meanings was randomised for each participant.

364   Finally, participants provided their demographics details and answered some questions

365   about their reading habits. These questions were used to maintain the impression that the

366   experiment was investigating general reading and comprehension and the responses were not

367   analysed.

368   Exactly 24 hours after the first session, participants were invited to take part in a short

369   24-hour follow-up to the experiment. Thirty-one participants completed the follow-up tests,

370   which they did an average of 24 hours and 1 minute (SD = 54 minutes, range = 22h26m-28h2m)

371   after the first session. The follow-up tests consisted of a repeat of the two tests from the first

372   session of the experiment in the same order.

373

374   **Results**

375   **Analysis procedure.**

376   Responses from the multiple-choice test were either coded as "1" for correct or "0" for

377   incorrect with regards to which word had been selected to match with the meaning. Responses

378   for the cued recall test were independently coded for accuracy by the experimenter and a

379   research assistant, blind to condition, as either "1" for correctly recalled meanings or "0" for

380   incorrect[1]. Responses were leniently coded as correct if at least one correct semantic feature was

381   recalled. Any ambiguous or partially correct responses were resolved on a case-by-case basis

382   through discussion. One item ("bruise") was excluded from the analyses for the cued recall

383   measure, as the percentage of participants who gave a correct response for that item in one of the

384   two learning conditions (incidental, 20.0%) was more than two standard deviations below the

385   grand mean for all items across both learning conditions ($M = 72.3\%$; $SD = 23.5$).

386   Data were analysed with logistic mixed effects models using the *lme4* package

387   (version1.1-7; Bates et al., 2015) and R statistical software (version 3.0.2; R Core Team, 2017).

388   Four separate models were created: one for each test measure comparing the accuracy between

389   day one and day two (including only the participants who completed both test sessions, $N = 31$),

390   and for each measure for all participants tested on day one only ($N = 40$). These latter analyses

391   aimed to verify that the data from this larger set of participants did not differ from the subset who

392   chose to complete both sessions.

393   Due to a potential effect of counterbalancing order, in that participants who completed

394   the intentional condition first may suspect that they would be tested on the words in the stories,

395   we included a factor for the order of the learning conditions in our models. The contrasts for the

396   fixed effects were defined using deviation coding for learning condition (incidental: -0.5,

397   intentional: 0.5), and the order of the learning conditions (first: -0.5, second: 0.5), with the

398   interaction coded by multiplying the contrasts for these two factors. The two models comparing

399   performance between day one and day two contained additional fixed effects for time (day one: -

400   0.5, day two: 0.5), and the interactions between time and learning condition, time and learning

---

[1] While each new meaning comprised three distinguishing semantic features, due to the natural integration of the new word meanings into the stories not all of the semantic features were mentioned at each occurrence. Because it was more difficult to control even exposure to all of the semantic features in the incidental condition (stories), we were concerned that a scaled accuracy measure of the number of semantic features recalled would be biased towards the intentional condition. We therefore decided to score recall responses with a binary accuracy score.

401 order, and the three-way interaction. Random effects structures were determined by identifying

402 the maximal model (Barr et al., 2013). This included by-participant and by-item random

403 intercepts, and by-participant and by-item random slopes for learning condition[2]. The models

404 comparing performance between day 1 and day 2 also included by-participant and by-item

405 random slopes for time and the interaction between time and learning condition. Where the

406 maximal model failed to converge, we simplified the models by removing the correlations

407 between the by-participant and by-item random slopes and random intercepts without removing

408 any of the random slopes, as recommended by Barr et al. (2013). Significance of the fixed effects

409 and interactions was assessed using likelihood ratio tests comparing the full model to models

410 with each fixed factor/interaction of interest removed in turn (but leaving in any interaction

411 involving a factor of interest that has been removed). Follow-up analyses were carried out in the

412 case of any significant interaction using the same method. The p-values for the simple effects

413 analyses were compared against a Bonferroni-corrected α of .025. All data and analysis scripts

414 for this experiment are available via the Open Science Framework (OSF; https://osf.io/k32tw).

415

416   **Cued recall of meanings.**

417

418   **[Figure 1.png]**

419 **Figure 1. Experiment 1. Mean percentage of correct responses on the cued recall test for**

420 **each learning condition, when tested on day one (immediately after learning) and 24 hours**

421 **later ($N = 31$).**

---

[2] Random slopes for the controlled factor of learning order were not included due to issues of model non-convergence. This factor was not of theoretical interest and the fixed effect for this factor was not significant in any of the models.

422   The LME analyses were carried out on the raw binary accuracy data, however mean percentage

423   accuracy data are displayed in the graphs for ease of interpretation. Error bars show standard

424   errors for the means, adjusted for the within-participant design (Cousineau, 2005).

425

426   The accuracy data for cued recall of the new meanings comparing performance between

427   day one and day two ($N = 31$; see Figure 1) showed a reasonably high level of accuracy for items

428   learned through both learning conditions. But accuracy was significantly higher overall for items

429   learned through the intentional learning condition (day one: 85.2%; day two: 84.9%) than those

430   learned under incidental conditions (day one: 62.1%; day two: 70.1%) [$\chi^2(1) = 14.32, p < .001$].

431   There was no significant main effect of time of test [$\chi^2(1) = 1.23, p = .268$], nor of learning order

432   [$\chi^2(1) = 0.83, p = .362$]. While accuracy improved slightly between day one and day two for

433   items learned through the stories and remained at a similar level for items learned through the

434   intentional condition, the interaction between learning condition and time was non-significant

435   [$\chi^2(1) = 1.57, p = .210$]. The interactions between learning condition and learning order [$\chi^2(1) =$

436    $2.16, p = .141$], learning order and time of test [$\chi^2(1) = 1.79, p = .181$], and the three-way

437   interaction [$\chi^2(1) = 0.83, p = .361$] were also non-significant.

438   The accuracy data for all participants on day one only ($N = 40$) showed a similar pattern:

439   significantly higher accuracy for items learned through the intentional training condition than the

440   incidental condition [$\chi^2(1) = 21.35, p < .001$], and no significant main effect of learning order

441   [$\chi^2(1) = 0.0007, p = .979$]. However, there was a significant interaction between learning

442   condition and learning order [$\chi^2(1) = 4.68, p = .030$]: accuracy was higher for items learned

443   through the stories when this had been the first condition in the experiment (70.0%) than when

444   this had been the second condition (47.5%), whilst the opposite was the case for items learned

445 through the intentional condition (first: 82.1%; second: 95.0%). This unexpected result could

446 have been driven by participants becoming somewhat fatigued by the second task and this

447 fatigue effect having greater impact on the story reading task, which took longer. However, this

448 does not perhaps explain why participants performance in the intentional condition was slightly

449 higher when it was the second condition in the experiment rather than the first condition.

450

451   **Multiple-choice meaning-to-word matching.**

452

453   **[Figure 2.png]**

454 **Figure 2. Experiment 1. Mean percentage of correct responses on the multiple-choice test**

455 **for each learning condition, when tested on day one (immediately after learning) and day**

456 **two (24 hours later; $N = 31$).**

457 Error bars show standard errors for the means, adjusted for the within-participant factor of

458 learning condition (Cousineau, 2005).

459

460   The results for accuracy on the multiple-choice meaning-to-word matching test

461 comparing results between day one and day two ($N = 31$) are shown in Figure 2. Accuracy was

462 high in both learning conditions, but was slightly higher for the intentional learning condition

463 (day one: 96.0%; day two: 87.9%) than for the incidental learning condition (day one: 83.9%;

464 day two: 83.1%), although this difference was non-significant [$\chi^2(1) = 3.66, p = .056$]. The main

465 effect of time was also non-significant [$\chi^2(1) = 3.81, p = .051$], as was the main effect of learning

466 order [$\chi^2(1) = 0.002, p = .966$]. Interestingly, the interaction between learning condition and time

467 was significant [$\chi^2(1) = 3.85, p = .05$]. The interaction between learning condition and learning

468   order was non-significant [$\chi^2(1) = 2.24$, $p = .135$], as was the interaction between time and

469   learning order [$\chi^2(1) = 0.01$, $p = .929$], and the three-way interaction [$\chi^2(1) = 0.48$, $p = .488$].

470   To follow up on the significant interaction between learning condition and time, two

471   simple effects analyses were carried out to determine the significance of time within each of the

472   two learning conditions separately. For the incidental learning condition there was no significant

473   effect of time [$\chi^2(1) = 0.10$, $p = .750$], indicating no forgetting between day one and day two.

474   However, there was a significant effect of time for the intentional learning condition [$\chi^2(1) =$

475   6.07, $p = .014$], which was slightly lower on day two than day one. (The p-values for these

476   simple effects analyses were compared against a Bonferroni-corrected α of .025.)

477   In the accuracy data for the multiple-choice test for all participants on day one only ($N =$

478   40), accuracy was again high overall for both conditions. Accuracy appeared slightly higher for

479   items learned through the intentional condition (95.7%) than for items learned through the

480   incidental condition (81.9%), but there was no significant main effect of learning condition [$\chi^2(1)$

481    $= 1.18$, $p = .277$]. There was also no significant main effect of learning order [$\chi^2(1) = 0.004$, $p =$

482    $.950$], and no significant interaction [$\chi^2(1) = 2.52$, $p = .113$].

483

484   **Discussion**

485   Experiment 1 aimed to determine how easily novel meanings for familiar words can be

486   acquired incidentally through story reading, as compared with a more intentional learning

487   procedure. The results showed that accuracy in cued recall of new word meanings was

488   significantly higher in the intentional learning condition than the incidental learning condition:

489   85.2% compared with 62.1% when measured immediately after training. The accuracy data for

490   the multiple-choice meaning-to-word matching test showed a similar pattern (96.0% for the

491   intentional condition and 83.9% for the incidental condition at the immediate test), although for

492   this measure the main effect of learning condition was non-significant. Furthermore, reading the

493   story took participants a significantly longer amount of time, so more was learned in a shorter

494   amount of time through the intentional training task.

495   These findings are broadly consistent with those of L2 vocabulary learning studies that

496   have found intentional learning to be more efficient than incidental learning of vocabulary

497   (Hulstijn, 1992; Peters et al., 2009). Although recall accuracy was higher for the intentional

498   learning condition, there was also a reasonably high level of acquisition of new meanings for

499   familiar words through the incidental learning condition. This is very similar to the results of the

500   study by Hulme et al. (2019), where accuracy in recalling new meanings for familiar words was

501   63.5% after eight exposures in an incidental learning context (accuracy in cued recall of word

502   forms was 69.2%). These results therefore support the findings of recent studies showing good

503   acquisition of L1 vocabulary from reading (Batterink & Neville, 2011; Godfroid et al., 2017;

504   Pellicer-Sánchez, 2016). An additional consideration is that we included multiple-choice

505   comprehension questions between pages of the stories in our incidental learning condition. While

506   these did not probe details of the new word meanings, it is possible that they could have

507   improved comprehension by increasing metacognitive awareness of the comprehension process,

508   similar to a "guided learning" scenario (Blything et al., 2020).

509   Interestingly, the multiple-choice test showed a significant interaction between learning

510   condition and time of test, although there was no such significant interaction in the recall

511   measure. Following up on the significant interaction in the multiple-choice measure, simple

512   effects analyses showed that there was significant forgetting after 24 hours of items learned in

513   the intentional condition (8% reduction in accuracy between the immediate test and delayed test),

514   but there was no forgetting of items learned incidentally through story reading. This absence of

515   significant forgetting after a 24-hour delay replicates the finding from Hulme et al. (2019) who

516   reported no significant forgetting on this paradigm after a seven-day delay. A possible

517   explanation for this lack of forgetting is that new word meanings learned in a more semantically

518   rich context, such as from stories, may be retained better. New word meanings encountered in

519   stories will contain additional contextual information relating to the narrative (e.g., characters'

520   thoughts and feelings), providing additional cues for participants to rely on for later retrieval.

521   The stories are also more interesting and engaging so may be more memorable for participants in

522   general. However, another possibility is that this difference is a function of initial learning level,

523   as initial performance was very high in the intentional condition and so had further to fall. A

524   further alternative explanation for this finding is that the lack of forgetting of items learned

525   incidentally through story reading could be due to memory reactivations during sleep. Memory

526   consolidation during sleep has been shown to be preferential towards initially weaker memory

527   traces and less supportive to memories that have already established robust representations

528   (Drosopoulos et al., 2007). Since recall and recognition accuracy was lower for items acquired

529   through incidental learning conditions, this is a compelling alternative explanation for why there

530   was less forgetting of items learned through story reading.

531   In addition, the additional retrieval practice for the test immediately after training may

532   have aided learning in the incidental, story reading condition. Further supporting this possibility,

533   items learned through the incidental condition showed a slight improvement (8% increase) in

534   cued recall accuracy between the test on day one and the test on day two. The second test task in

535   the immediate test session (multiple-choice meaning-to-word matching) may therefore have

536   boosted learning, manifesting as improved cued recall at the delayed test. The potential

537   involvement of a testing effect (Roediger & Karpicke, 2006a) in long-term retention of new

538   meanings for familiar words will be examined in detail in Experiment 2 and Experiment 3.

539

540   **Experiment 2: The testing effect in incidental and intentional learning**

541   The aim of Experiment 2 was to investigate whether 24-hour retention of novel word

542   meanings that were learned through story reading can be boosted by introducing a test of

543   participants' knowledge immediately after training. As with Experiment 1, performance will be

544   compared to a more conventional explicit learning baseline (the same intentional training

545   procedure as in Experiment 1). Participants were tested immediately after training on half of the

546   items they saw in each learning condition, and then given a surprise test 24 hours later in which

547   they were tested on the other half of the items for the first time, as well as being retested on items

548   that had been tested the previous day. As for Experiment 1, both test sessions consisted of cued

549   recall of the new meanings, followed by a multiple-choice meaning-to-word matching test.

550   We predicted that there would be better long-term retention for items that were tested

551   immediately after training, compared with those that were not tested, for both the incidental and

552   intentional conditions. There was no specific prediction as to whether the magnitude of the

553   testing effect would differ for the different learning conditions. Additionally, we expected to

554   replicate the findings from Experiment 1 that retention would be better overall for novel

555   meanings learned through intentional conditions, but that there would be less forgetting under

556   incidental conditions.

557   Experiment 2 was preregistered through the Open Science Framework; the preregistration

558   can be retrieved from https://osf.io/e5zmk (Hulme & Rodd, 2016, November 4). Any deviations

559   from the preregistration are noted in the Method and Results sections for this experiment. The

560   materials, data, and analysis scripts for Experiment 2 can be found on the Open Science

561   Framework (OSF; https://osf.io/upmnr).

562

563   **Method**

564   **Participants.**

565   We aimed to recruit 96 participants for Experiment 2 in which participants were trained

566   on eight items (two items per cell: four per learning condition, and four per test type) in one of

567   16 experiment versions (six participants per version). The sample size was established in

568   consideration of the study by Hulme et al. (2019) and Experiment 1. As this experiment has the

569   additional independent variable of test type, a larger sample size was used in Experiment 2 than

570   Experiment 1 to achieve comparable power with a smaller number of items per cell.

571   Ninety-nine participants were included in the experiment (age: $M = 32.31$ years, $SD =$

572   8.14; 56 female); we accidentally over-recruited by three participants when pseudorandomly

573   assigning participants to the experiment versions and kept these participants. Participants were

574   recruited in the same way as for Experiment 1. They gave their informed consent before taking

575   part (by means of ticking boxes in the online consent form) and were paid for their participation

576   at the end of each session (£6 for session one and £2 for session two). The UCL Experimental

577   Psychology Ethics Committee granted ethical approval for the research (Ref: EP/2017/009).

578   An additional 36 participants took part in the first session but did not complete session

579   two by the deadline (within six hours of receiving the invitation for the delayed test) and were

580   excluded. (This additional data exclusion was used for Experiments 2 and 3 to ensure sufficient

581   power to examine the testing effect within-participants, for which a complete set of immediate

582   and delayed test data was critical.) A further twenty-one participants were excluded due to

583   getting more than one of the multiple-choice comprehension questions wrong when reading the

584   stories (see Experiment 1 Procedure), and two further participants were excluded for attempting

585   the experiment more than once. Finally, five participants were excluded for being outliers in their

586   mean reading speed (faster than 543.4 words per minute, 2 *SD* above the mean). Excluded

587   participants were replaced during recruitment.

588

589   **Materials.**

590   The stimuli for the present experiment were identical to those used in Experiment 1.

591

592   **Design.**

593   The experiment used a within-participants and within-items design, with two independent

594   variables: learning condition (two levels: incidental and intentional) and test type (three levels:

595   immediate test (tested in the first session), delayed test (tested for the first time in the second

596   session), and delayed retest (tested for the second time in the second session)). The dependent

597   variables were accuracy in cued recall of meanings and multiple-choice meaning-to-word

598   matching.

599   There were sixteen versions of the experiment to ensure the items were seen an even

600   number of times in each condition, with the order of the learning conditions counterbalanced

601   across participants. Participants were pseudorandomly assigned to one of the sixteen versions of

602   the experiment. As in Experiment 1, each participant was trained on half the total number of

603   stimuli (eight items; four in each learning condition), with the items in each condition and the

604   order of the learning conditions counterbalanced across participants. Further counterbalancing

605   accounted for which stimulus items were or were not tested immediately following training

606   across participants.

607

608   **Procedure.**

609   The procedure for the incidental and intentional learning conditions was identical to that

610   of Experiment 1, and immediately following training participants completed the same Mill Hill

611   vocabulary test (Mill Hill Vocabulary Test, Set A: Multiple Choice: Buckner et al., 1996; Raven

612   et al., 1998) as a filler task. Participants were then immediately tested on half of the items that

613   they had been trained on through the incidental and intentional learning conditions (four items,

614   two trained through each training method). The tests were the same as for Experiment 1: cued

615   recall of meanings followed by multiple-choice meaning-to-word matching. The items were

616   tested in a random order in each of the two tests, with no feedback given to participants. In the

617   multiple-choice test only the four words that a participant was being immediately tested on

618   appeared as the four alternative responses to choose from for each test item; the order of these

619   was also randomised for each test item.

620   Exactly 24 hours after the first session of the experiment had been made available,

621   participants were asked to take part in the second session: the delayed test. Participants were not

622   aware beforehand that they would be asked to complete this test to discourage them from

623   rehearsing and intentionally retaining information about the novel word meanings. As such,

624   unfortunately 36 participants did not return to complete session two; 99 participants were

625   therefore included in the analysis. Participants began the delayed test an average of 24 hours and

626   25 minutes ($SD = 57$ mins, range $= 22h45m-27h21m$) after the training session. The tests were

627   the same as those that had been used for the immediate test in the same order. This time

628   participants were tested on all of the stimuli that they had been trained on (eight items). The

629   order of presentation of the items in each of the two tests was again randomised for each

630   participant, and for the multiple-choice test the order of the eight stimulus words to choose from

631   was again randomised for each test item.

632

633   **Results**

634   **Analysis procedure.**

635   Responses for the cued recall test and multiple-choice test were coded for accuracy ("1"

636   for correct and "0" for incorrect) in the same way as for Experiment 1.

637   Upon completion of the experiment, we realised that test type was confounded with

638   differing test difficulty between the immediate test and the two delayed test types for the

639   multiple-choice meaning-to-word matching measure. This was because in the immediate test

640   participants could choose from four alternative words to pair with the appropriate meaning on

641   each trial (as they were only tested on half the total items they were trained on: four items, two

642   trained through each training condition). In contrast, in the delayed test participants had to

643   choose from eight alternatives (as they were tested on all eight of the items they had been trained

644   on). The results from the immediate multiple-choice test are therefore not comparable to the

645   results from the two delayed test types, so the analysis for this measure was only carried out on

646   the subset of results for the two delayed test types. This is a deviation from the analysis plan

647   outlined in the preregistration of this experiment. The analysis of the cued recall measure was

648   carried out according to the preregistration.

649   The data were analysed, as in the previous experiment, using logistic mixed effects

650   models with the *lme4* package (version 1.1-12; Bates et al., 2015) and R statistical software

651   (version 3.3.2, R Core Team, 2017), with two separate models for the analysis of data from the

652   two measures. The model used to analyse the cued recall data contained three factors: test type

653   (three levels: immediate, delayed first test, delayed second test), learning condition (two levels:

654   incidental, intentional), and order of the learning conditions in the experiment (two levels: first,

655   second). The contrasts for the fixed effect of test type were defined using Helmert coding, with

656   one contrast comparing the immediate test to the two delayed tests combined (immediate: 0.67,

657   delayed first test: -0.33, delayed second test: -0.33), and a second comparing the two delayed test

658   types to each other (immediate: 0, delayed first test: -0.5, delayed second test: 0.5). Deviation

659   coding was used to specify the contrasts for the fixed effects of learning condition (incidental: -

660   0.5, intentional: 0.5) and learning order (first: -0.5, second: 0.5).

661   The model used to analyse the multiple-choice data also had three factors: test type (2

662   levels: delayed first test, delayed second test), learning condition (2 levels: incidental,

663   intentional), and learning order (2 levels: first, second). The contrasts were specified using

664   deviation coding for the fixed effects of test type (delayed first test: -0.5, delayed second test:

665   0.5), learning condition (incidental: -0.5, intentional: 0.5), and learning order (first: -0.5, second:

666   0.5). The procedure for determining the appropriate random effects structure and significance of

667   the fixed effects/interactions was the same as described for Experiment 1. The model used to

668   analyse the multiple-choice data used the maximal random effects structure; the one for the cued

669   recall measure was simplified by removing the correlations between the by-participant and by-

670   item random slopes and random intercepts (as recommended by Barr et al., 2013).

671   Following on from the main analysis for the cued recall measure, firstly three pairwise

672   comparisons (with Bonferroni adjustment for multiple comparisons, $\alpha = .017$) were carried out to

673   compare the different levels of test type to each other. This was done by taking a subset of the

674   data for each pair of levels of test type and creating a model for each containing the same fixed

675   and random effects as the model used for the main analysis, although the contrast for test type

676   was coded using deviation coding (immediate: 0.5, delayed first test: -0.5; immediate: 0.5,

677   delayed second test: -0.5; delayed first test: -0.5, delayed second test: 0.5). Secondly, further

678   follow-up pairwise comparisons (with Bonferroni correction for multiple comparisons, $\alpha = .017$)

679   were made for the three 2x2 interactions between the pairs of test types and the two learning

680   conditions to determine whether the difference between any two test types differed between the

681   two learning conditions. Finally, six simple effects pairwise comparisons (with Bonferroni

682   adjustment for multiple comparisons, $\alpha = .008$) were run to test for any significant differences

683   between the different test types within the two learning conditions. This was done by taking

684   further subsets of the data for the pairs of levels of test type separately for the incidental and

685   intentional learning conditions and creating models with only fixed effects for test type, learning

686   order, and the interaction (with random effects for test type for participants and items).

687   The only follow-up analyses carried out for the multiple-choice test were two simple

688   effects pairwise comparisons (with Bonferroni correction for multiple comparisons, $\alpha = .025$).

689   This was done in the same way as the simple effects analyses for the cued recall measure. All

690   data and analysis scripts for this experiment are available via the Open Science Framework

691   (OSF; https://osf.io/upmnr).

692

693   **Cued recall of meanings.**

694   Data for the cued recall test (Figure 3) showed that accuracy was significantly higher for

695   items trained through the intentional than the incidental learning condition [$\chi^2(1) = 34.83$, $p <$

696   .001]. There was also a significant main effect of test type [$\chi^2(2) = 25.78$, $p < .001$]; the main

697   effect of learning order was non-significant [$\chi^2(1) = 2.50$, $p = .114$]. There was a significant

698   interaction between learning condition and test type [$\chi^2(2) = 13.86$, $p < .001$], but the interaction

699   between learning condition and learning order was not significant [$\chi^2(1) = 0.09$, $p = .760$]. There

700   was an unexpected significant interaction between test type and learning order [$\chi^2(2) = 9.24$, $p =$

701   .010]; the three-way interaction was not significant [$\chi^2(2) = 3.23$, $p = .199$]. (Results for all the

702   analyses for the cued recall test are also presented in Table 1 for clarity of the different levels of

703   follow-up analysis for this measure.)

704

705   **[Figure 3.png]**

706   **Figure 3. Experiment 2. Mean percentage of correct responses in the cued recall test for**

707   **each learning condition and for the three different test types in the experiment.**

708   Error bars show standard error of the means adjusted for the within-participant design

709   (Cousineau, 2005).

710

711   **[Table 1.docx]**

712   **Table 1. Results for the linear mixed effects model analyses for the cued recall test in**

713   **Experiment 2 showing the main analyses and the different levels of follow-up analyses.** Test

714   results reported are for likelihood ratio tests comparing the full model to models with each fixed

715   factor/interaction of interest removed in turn.

716

717   To further investigate the significant main effect of test type, three pairwise comparisons

718   between the different levels of test type were carried out. The results revealed that there was a

719   significant effect of overnight forgetting (difference between the delayed 1st and immediate tests)

720 $[\chi^2(1) = 19.99, p < .001]$, with better recall of new meanings when tested immediately than when

721 tested for the first time after a delay. There was also a significant testing effect (difference

722 between the delayed 2nd and delayed 1st tests) $[\chi^2(1) = 18.83, p < .001]$, with higher recall

723 accuracy for items that were being tested for the second time than for those being tested for the

724 first time after the delay. However, there was no significant difference in cued recall accuracy

725 between the immediate and delayed 2nd tests $[\chi^2(1) = 0.89, p = .345; \alpha = .017]$, suggesting these

726 items were protected against forgetting.

727   To further investigate the significant interaction between learning condition and test type,

728 the second set of follow-up analyses were pairwise comparisons for the three 2x2 interactions

729 between the pairs of test types and the two learning conditions. Results showed that there was a

730 significant interaction between learning condition and the difference between the immediate and

731 delayed 2nd tests $[\chi^2(1) = 16.24, p < .001]$. Items learned incidentally from stories showed some

732 improvement between the immediate test and the retest on day two, while items learned through

733 the intentional learning condition showed a small amount of forgetting. There was no significant

734 interaction between learning condition and the difference between either the immediate and

735 delayed 1st tests $[\chi^2(1) = 2.29, p = .130]$ or between the delayed 1st and delayed 2nd tests $[\chi^2(1) =$

736 $2.97, p = .085; \alpha = .017]$.

737   In the final set of follow-up analyses, six simple effects pairwise comparisons were run to

738 test for any significant differences between the different test types within the two learning

739 conditions. The results revealed that for the incidental learning condition there was no significant

740 difference in recall accuracy for items tested for the first time after the delay than for items tested

741 immediately after training at the corrected level $[\chi^2(1) = 5.88, p = .015]$, that is no significant

742 forgetting. There was significantly better cued recall accuracy for items tested for the second

743 time after the delay than those tested for the first time [$\chi^2(1) = 15.27, p < .001$]: a significant

744 testing effect. There was also significantly better recall of new meanings tested for the second

745 time after the delay than the immediate test [$\chi^2(1) = 14.59, p < .001$]. For the intentional learning

746 condition, there was significantly lower recall accuracy for items tested for the first time after the

747 delay than those tested immediately [$\chi^2(1) = 18.27, p < .001$], showing significant forgetting.

748 There was better recall accuracy for items tested for the second time after the delay than those

749 tested for the first time [$\chi^2(1) = 8.39, p = .004$]: a significant testing effect. There was no

750 significant difference (at the corrected level) between items tested for the second time following

751 the delay and when tested immediately after training [$\chi^2(1) = 4.25, p = .039; \alpha = .008$].

752   Additionally, although not specified in the preregistration, exploratory follow-up analyses

753 were carried out to examine the nature of the unexpected interaction between test type and

754 learning order. These were three pairwise comparisons of the 2x2 interactions between the pairs

755 of test types and the two learning orders (first or second position in the experiment). Results

756 revealed a significant interaction between position and the difference between the delayed 1st and

757 delayed 2nd tests [$\chi^2(1) = 7.12, p = .008$]. Items appeared to be recalled better at the delayed 1st

758 test when they had been presented in the first condition in the training session, whereas items

759 were recalled better at the delayed 2nd test when they had been trained in the second condition.

760 There was no significant interaction between the immediate and delayed 2nd tests [$\chi^2(1) = 0.12, p$

761 $= .729$], nor between the immediate and delayed 1st tests [$\chi^2(1) = 5.15, p = .023$] at the

762 Bonferroni-corrected level ($\alpha = .017$). However, further follow-up analyses of the simple effects

763 pairwise comparisons of learning order within the delayed 1st and delayed 2nd test types were

764 both non-significant at the Bonferroni-corrected level (both $p > .025$).

765

766   **Multiple-choice meaning-to-word matching.**

767   The analysis of the data for the multiple-choice meaning-to-word matching test (Figure 4)

768   was only carried out on the subset of results for the two delayed test types. Overall accuracy was

769   very high, and it was significantly higher for items trained intentionally than incidentally across

770   all test types [$\chi^2(1) = 8.44$, $p = .004$]. The main effect of test type was also significant [$\chi^2(1) =$

771   6.71, $p = .010$], with slightly greater accuracy for items that had been tested previously than for

772   those that had not been; there was no significant main effect of learning order [$\chi^2(1) = 0.32$, $p =$

773    $.569$]. The interaction between learning condition and test type was not significant [$\chi^2(1) = 0.61$,

774    $p = .435$], nor was the interaction between learning condition and learning order [$\chi^2(1) = 1.23$, p

775    $= .268$], nor the interaction between test type and learning order [$\chi^2(1) = 1.32$, $p = .251$]. The

776   three-way interaction was also not significant [$\chi^2(1) = 0.06$, $p = .810$].

777

778   **[Figure 4.png]**

779   **Figure 4. Experiment 2. Mean percentage of correct responses in the multiple-choice test**

780   **for each learning condition and for the three different test types in the experiment.**

781   Note that the results from the immediate test (with lighter shading) are not comparable to those

782   from the two delayed test types due to an underlying difference in test difficulty. Error bars show

783   standard error of the means adjusted for the within-participant design (Cousineau, 2005).

784

785   Following on from the main analysis, two simple effects pairwise comparisons tested for

786   any significant differences between the different test types within the two learning conditions. In

787   the incidental learning condition, the difference between items tested for the second time

788   following the delay and those tested for the first time was non-significant at the corrected level

789    $[\chi^2(1) = 4.62, p = .032]$. For the intentional learning condition, there was also no significant

790   difference in accuracy between items tested for the first or second time after the delay $[\chi^2(1) =$

791    $3.53, p = .060; \alpha = .025]$.

792

793   **Discussion**

794   The aim of Experiment 2 was to examine whether testing memory immediately after

795   training enhances long-term retention of new word meanings acquired through incidental and

796   intentional learning conditions, and to see whether any testing effect differs depending on the

797   learning conditions. As in Experiment 1, new word meanings were learned better overall through

798   intentional learning conditions than through incidental learning conditions. Accuracy in both the

799   cued recall and multiple-choice tests was significantly higher for items trained through the

800   intentional learning condition

801   Cued recall accuracy was also higher overall immediately after training than when items

802   were tested for the first time after 24 hours. This demonstrates some overnight forgetting of the

803   new meanings for the words in the absence of an intervening test. Furthermore, there was

804   numerically but not significantly (at the corrected level) more forgetting of items trained through

805   the intentional learning condition than those learned through a story, both with and without prior

806   retrieval practice. This is in line with the findings of Experiment 1, where accuracy on the

807   multiple-choice test was lower after 24 hours for items learned through the intentional condition,

808   but not for items learned through the stories.

809   Critically, both the cued recall and multiple-choice tests revealed an overall testing effect:

810   new word meanings were recalled and recognised significantly better after 24 hours when they

811   had been tested immediately after training than when they were being tested for the first time. As

812   predicted, in the cued recall measure this main effect of testing was also significant in the simple

813   effects analyses that looked at incidental and intentional learning separately. This is in line with

814   studies that have found a benefit of prior retrieval on learning information from different

815   contexts, such as list of foreign language vocabulary words and their translations (van den Broek

816   et al., 2013) and information from prose passages (Roediger & Karpicke, 2006b).

817   The lack of difference in cued recall accuracy between performance on the immediate

818   test and the delayed test for items being tested for the second time suggests that the retrieval

819   practice protected these items against forgetting. The testing effect seen in the present study

820   therefore at least partly explains why participants in Experiment 1 and those in the study by

821   Hulme et al. (2019) showed such good retention after one day and one week respectively. In the

822   present experiment both the cued recall and multiple-choice tests were administered to all

823   participants at both time points. It is therefore unclear whether either of these tests on its own

824   would produce a similar testing effect, or if the combination of the two was important for

825   boosting long-term retention. It also remains to be seen whether one of these test types is better

826   than the other for enhancing retention of new word meanings learned incidentally through

827   reading.

828   In sum, Experiment 2 demonstrated that testing memory of new meanings for familiar

829   words benefits their future retention. This was the case for recalling word meanings learned

830   either incidentally through story reading or through an intentional learning condition. As in

831   Experiment 1, participants learned vocabulary more efficiently through the intentional learning

832   condition, but performance for both learning conditions was good. There was a trend in the data

833   towards less forgetting of items trained incidentally through the stories, and a trend suggesting a

834   larger testing effect for incidentally-trained items. However, these interactions were non-

835   significant; further research is warranted to investigate whether incidentally-trained items in

836   particular could benefit from the additional learning opportunity afforded by the immediate test.

837   Either the immediate cued recall or multiple-choice test, or indeed a combination of the two, may

838   have produced the observed testing effect; Experiment 3 will investigate which of these test

839   methods could be more beneficial for retention. The results of this experiment will guide future

840   development of real-world intervention studies aimed to boost vocabulary learning from story

841   reading.

842

843   **Experiment 3: Immediate test method**

844   The testing effect has been observed in studies using various methods of immediate test,

845   most usually with cued recall (e.g., Karpicke & Smith, 2012), but also with other methods such

846   as multiple-choice (e.g., Roediger & Marsh, 2005). There are several possibilities as to why

847   certain methods of immediate testing may be more beneficial for future retention. The retrieval

848   effort hypothesis states that testing is more helpful for long-term retention when it is more

849   effortful (Pyc & Rawson, 2009). For example, in a study in which young adults learned the

850   meanings of novel L1 vocabulary words, Karpicke and Roediger (2007) showed that increasing

851   retrieval difficulty by increasing the delay between initial study and initial testing led to better

852   long-term retention than when initial retrieval effort was lower. Tests of productive vocabulary

853   knowledge, such as cued recall of word meanings, are more difficult than recognition tests in

854   which word meanings are supplied (Pellicer-Sánchez, 2016); therefore an immediate cued recall

855   test may be more advantageous for future retention than a multiple-choice test. Indeed, findings

856   from many studies (for a review see Rowland, 2014) suggest effortful processing to be an

857   important attribute of the testing effect.

858   Conversely, immediate testing may be particularly beneficial when it assists with

859   restructuring learned information into a format that is more helpful for long-term retention.

860   Multiple-choice recognition tests may aid retention due to response choices cueing the retrieval

861   of marginal knowledge that may otherwise not be easily accessible (Marsh et al., 2007). They

862   may also provide an opportunity for additional learning of some items through the process of

863   elimination of foils (Marsh et al., 2007) even in the absence of feedback on response choice.

864   However, foil answers in multiple-choice tests may also lead to learning of incorrect information

865   (Butler et al., 2006; Marsh et al., 2007; Roediger & Marsh, 2005).

866   Some studies have directly compared the effects of immediate cued recall and multiple-

867   choice tests on long-term retention. Duchastel (1981) investigated secondary school students'

868   retention of a prose passage following immediate testing with either a short-answer test (akin to

869   cued recall), a multiple-choice test, or a free recall test. Long-term retention (measured by a

870   delayed cued recall test) was better for those who had the immediate short-answer test, but no

871   testing effect was observed for the other two groups. However, Duchastel (1981) found no

872   testing effect for any group on the delayed free recall test, and the delayed cued recall measure

873   was very similar to the immediate test for the short-answer test group. More recently, Nakata

874   (2016) compared retrieval methods including cued recall and multiple-choice recognition in a

875   study of paired-associate learning of novel L2 words. Recall was found to be most beneficial for

876   acquiring novel words' orthography (spelling), whereas recognition was more beneficial

877   otherwise (Nakata, 2016).

878   One concern is that information learned with the help of retrieval practice could be

879   relatively inflexible and constrained, and may therefore not transfer to different delayed tests.

880   Tran et al. (2014) and others have found that retrieval practice may not benefit later tests that

881   require making deductive inferences about the learned information. Furthermore, Hogan and

882   Kintsch (1971) found that immediate test methods that provide further exposure (i.e., recognition

883   tests) were more beneficial than free recall for recognition two days later, whereas both free

884   recall and recognition boosted performance on delayed free recall.

885   However, the degree to which different methods of immediate test aid future retention

886   can also differ depending on factors such as the provision of feedback. Kang et al. (2007) found

887   that participants who had an immediate multiple-choice test performed better on delayed

888   multiple-choice and short-answer tests than participants who had an immediate short-answer test

889   (Experiment 1; Kang et al., 2007). However, in a second experiment where feedback was

890   provided on initial test performance (Experiment 2; Kang et al., 2007), the group with the

891   immediate short-answer test performed better on the delayed tests than those whose immediate

892   test had been multiple-choice, supporting the retrieval effort hypothesis (Pyc & Rawson, 2009).

893   Other studies have also found that the testing effect can transfer across different test methods

894   (Butler, 2010; McDaniel, Anderson, et al., 2007; Rohrer et al., 2010), with cued recall usually

895   found to be more beneficial for long-term retention than recognition tests.

896   The aim of Experiment 3 was to investigate the impact of immediate test method (cued

897   recall vs. multiple-choice meaning-to-word matching) on retention of new word meanings

898   learned through stories. A secondary aim was to rule out the possibility that the testing effect in

899   Experiment 2 was simply a practice effect due to having previously completed the same test, and

900   that retrieval practice can generalise to a different delayed test. We predicted that the results

901   would replicate the key finding from Experiment 2 of better long-term retention for items tested

902   immediately after training (regardless of testing method) than items not tested previously.

903   Additionally, we predicted that cued recall would be more beneficial for long-term retention of

904   new word meanings than multiple-choice meaning-to-word matching as, according to the

905   retrieval effort hypothesis (Pyc & Rawson, 2009), production tests that require more effortful

906   retrieval than recognition tests (Roediger & Butler, 2011; Rowland, 2014) are more helpful for

907   retention. However, it was also possible that the multiple-choice meaning-to-word matching test

908   could result in better subsequent retention as this test provides participants with additional cues

909   that provide an additional learning opportunity (Marsh et al., 2007). Finally, it was possible that

910   the testing effect would not transfer across test tasks (Hogan & Kintsch, 1971; Tran et al., 2014),

911   and so the benefit of each method of immediate test would only be seen for the delayed test of

912   the same type, in which case it could be characterised as more of a practice effect.

913   Experiment 3 was preregistered through the Open Science Framework; the preregistration

914   is available at https://osf.io/c59tz (Hulme & Rodd, 2017, June 23). Any deviations from the

915   preregistration are noted in the Method and Results sections for this experiment. The materials,

916   data, and analysis scripts for Experiment 3 can be found on the Open Science Framework (OSF;

917   https://osf.io/eh2c6).

918

919   **Method**

920   **Participants.**

921   We aimed to recruit 96 participants for Experiment 3 in which participants were trained

922   on eight items (four per testing condition) in one of two groups (who had different immediate

923   tests), with eight experiment versions (12 participants per version). The sample size was

924   established in consideration of the previous experiments. While there were more items per cell

925   per participant in Experiment 3 than Experiment 2, the additional between-participants factor of

926   test method meant that a similarly large sample was required to achieve comparable power in

927   this experiment.

928   Ninety-eight participants were included in the experiment (age: $M = 33.7$ years, $SD = 8.0$;

929   64 female); we over-recruited by two participants when pseudorandomly assigning participants

930   to the experiment versions and kept these participants. Participants were recruited in the same

931   way as for Experiments 1 and 2. They gave their informed consent before taking part (by means

932   of ticking boxes in the online consent form) and were paid for their participation at the end of

933   each session (£4 for session one and £2 for session two). The UCL Experimental Psychology

934   Ethics Committee granted ethical approval for the research (Ref: EP/2017/009).

935   An additional 18 participants took part in the first session but did not complete session

936   two by the deadline (within six hours of receiving the invitation for the delayed test) and were

937   excluded. A further thirty-five participants were excluded due to getting more than one of the

938   multiple-choice comprehension questions wrong in either of the stories they read (see

939   Experiment 1 Procedure). Seven further participants were excluded due to a technical issue

940   during data collection, and two participants were excluded for being outliers in their mean

941   reading speed (faster than 806.2 words per minute, 2 SD above the mean). Excluded participants

942   were replaced during recruitment.

943

944   **Materials.**

945   The stimuli for Experiment 3 were identical to those used in the previous experiments.

946   One additional different paraphrased version of each of the definition sentences was created so

947   that a differently worded definition would be presented in the immediate and delayed multiple-

948 choice meaning-to-word matching tests to counteract any direct practice effects (see Table S3 for

949 the additional paraphrased definitions: https://osf.io/tnb94).

950

951 **Design.**

952 The experiment used a mixed design with two independent variables: immediate test

953 method (two levels: cued recall and meaning-to-word matching) was manipulated between

954 participants, and whether items were or were not previously tested (two levels: not previously

955 tested versus previously tested was manipulated within subjects. The dependent variables were

956 accuracy in cued recall of the new meanings and multiple-choice meaning-to-word matching,

957 measured at the delayed test time point.

958 There were eight versions of the experiment to ensure the stimuli were seen an even

959 number of times in each condition across participants. Participants were pseudorandomly

960 assigned to one of the eight versions of the experiment. As in Experiments 1 and 2, each

961 participant was trained on half the total number of stimuli (eight items per participant; two

962 separate stories). For the key factor of immediate test method, half of the participants ($N = 49$)

963 were given a cued recall test of half of their items (four items) immediately after training, and the

964 other half of the participants ($N = 48$) were given a multiple-choice meaning-to-word matching

965 test of half of their items immediately after training. The items that were or were not tested

966 immediately following training were also counterbalanced across participants.

967

968 **Procedure.**

969 The first session of the experiment began with the incidental training procedure.

970 Participants first read one of the short stories; the procedure for this was identical to that of the

971   previous experiments. They were then asked to rate how enjoyable and clear they found the

972   story, and answer some questions about their subjective reading style, which took around two

973   minutes—this served as a brief interval between the two stories. Participants then read a second

974   story. Immediately following training participants completed the same Mill Hill vocabulary test

975   as used in the previous experiments (Mill Hill Vocabulary Test, Set A: Multiple Choice: Buckner

976   et al., 1996; Raven et al., 1998) as a filler task. Participants were then given an immediate test of

977   half the items they had been trained on (four items, two from each story), which was either a

978   cued recall test or a multiple-choice meaning-to-word matching test. The items were tested in a

979   randomised order in both of the test tasks, with no feedback given to participants. In the

980   multiple-choice meaning-to-word matching test, only the four words that a participant was being

981   immediately tested on appeared as the four alternative responses to choose from for each test

982   item; the order of these was also randomised for each test item.

983   Exactly 24 hours after the first session of the experiment had been made available, the

984   participants were asked to take part in the second session of the experiment: the surprise delayed

985   test. Again participants were not aware of this delayed test beforehand; unfortunately 18

986   participants did not return to complete session two and were replaced during data collection.

987   Participants completed the delayed test an average of 24 hours and 31 minutes ($SD = 57$ mins;

988   range = 22h40m-27h25m) after the training session. The tests used for the delayed test session

989   were the same as for the immediate test, but this time participants completed both tests: cued

990   recall of meanings followed by multiple-choice meaning-to-word matching. At the delayed test

991   participants were tested on all of the stimuli that they had been trained on (eight items). The

992   order of presentation of the items in each of the two tests was again randomised for each

993   participant. For the multiple-choice test, different paraphrased versions of the definition

994   sentences were used to those that had appeared in the immediate test, and the order of the eight

995   words to choose from was randomised for each test item.

996

997   **Results**

998   **Analysis procedure.**

999   Responses on the cued recall and multiple-choice tests were coded for accuracy in the

1000   same way as for the previous experiments. The data were again analysed using logistic mixed

1001   effects models with the *lme4* package (version 1.1-13; Bates et al., 2015) and R statistical

1002   software (version 3.3.3; R Core Team, 2017). Two models were created to analyse the results of

1003   the two delayed tests of cued recall and multiple-choice meaning-to-word matching separately.

1004   Contrasts for the fixed effects were defined using deviation coding for whether items were or

1005   were not immediately tested (not previously tested: -0.5, previously tested: 0.5), and the

1006   immediate test method (cued recall: -0.5, multiple-choice: 0.5), with the interaction coded by

1007   multiplying the contracts for these two factors. The model for the cued recall data used the

1008   maximal random effects structure, with by-participant and by-item random intercepts, a by-

1009   participants random slope for whether items were or were not previously tested, and by-items

1010   random slopes for whether items were or were not previously tested, method of immediate test,

1011   and the interaction. The model for the multiple-choice measure was simplified by removing the

1012   correlations between the random slopes and random intercepts (as recommended by Barr et al.,

1013   2013).

1014   Following on from the main analysis, simple effects analyses were carried out to

1015   determine significance of whether items had or had not been immediately tested within each of

1016   the two immediate test methods separately. All data and analysis scripts for this experiment are

1017   available via the Open Science Framework (https://osf.io/eh2c6).

1018

1019   **Cued recall of meanings.**

1020   The accuracy data for cued recall (measured in the delayed test session) are shown in

1021   Figure 5. Cued recall performance was low overall when items had not been tested immediately

1022   after training and was similar for the immediate cued recall group (26.5%) and immediate

1023   multiple-choice meaning-to-word matching group (25.5%): there was no significant main effect

1024   of immediate test method [$\chi^2(1) = 0.47, p = .491$]. This is reassuring as it shows that the two

1025   groups of participants (who had different immediate test methods) performed similarly overall.

1026   There was a significant main effect of whether items were or were not immediately tested [$\chi^2(1)$

1027 $= 23.73, p < .001$]: performance was much higher when items had been tested immediately after

1028   training. Although accuracy appeared higher for the immediate multiple-choice group (58.3%)

1029   than for the immediate cued recall group (49.5%) when items had been previously tested, the

1030   interaction was non-significant [$\chi^2(1) = 3.18, p = .074$]. The planned simple effects follow-up

1031   analysis showed that there was a significant effect of whether items were or were not

1032   immediately tested within the immediate cued recall group [$\chi^2(1) = 8.25, p = .004$], and also

1033   within the multiple-choice meaning-to-word matching group [$\chi^2(1) = 25.10, p < .001; \alpha = .025$].

1034

1035   **[Figure 5.png]**

1036   **Figure 5. Experiment 3. Mean percentage of correct responses in the cued recall test**

1037   **measured at the delayed test for participants whose immediate test was cued recall, and for**

1038 **those whose immediate test was multiple-choice when items were or were not previously**

1039 **tested.**

1040 Error bars show standard error of the means adjusted for the within-participants factor of whether

1041 items were or were not previously tested (Cousineau, 2005).

1042

1043   **Multiple-choice meaning-to-word matching.**

1044   The accuracy data for the multiple-choice meaning-to-word matching test (measured in

1045 the delayed test session) are shown in Figure 6. Performance on this test was much higher overall

1046 than on the cued recall test. Accuracy was significantly lower when items had not been tested

1047 immediately after training [$\chi^2(1) = 14.54, p < .001$], and there was no significant main effect of

1048 immediate test method [$\chi^2(1) = 2.38, p = .123$]. Concerning a possible interaction between

1049 whether items were or were not immediately tested and the immediate test method, the simple

1050 means for accuracy were similar for the immediate cued recall group (60.5%) and the immediate

1051 multiple-choice meaning-to-word matching group (63.5%) when there was no previous test. This

1052 is again reassuring as is shows that the two groups of participants (who had different immediate

1053 test methods) performed similarly overall. Although the simple means for accuracy on items that

1054 had been tested previously appeared to be higher for the immediate multiple-choice group

1055 (78.6%) than for the immediate cued recall group (67.5%), the interaction was non-significant

1056 [$\chi^2(1) = 2.48, p = .116$]. The planned simple effects follow-up analysis showed that the effect of

1057 whether items were or were not immediately tested was non-significant within the immediate

1058 cued recall group [$\chi^2(1) = 3.31, p = .069$]. The effect was, however, significant within the

1059 multiple-choice meaning-to-word matching group [$\chi^2(1) = 10.76, p = .001; \alpha = .025$].

1060

1061 **[Figure 6.png]**

1062 **Figure 6. Experiment 3. Mean percentage of correct responses in the multiple-choice test**

1063 **measured at the delayed test for participants whose immediate test was cued recall, and for**

1064 **those whose immediate test was multiple-choice, when items were or were not previously**

1065 **tested.**

1066 Error bars show standard error of the means adjusted for the within-participants factor of whether

1067 items were or were not previously tested (Cousineau, 2005).

1068

1069 **Discussion**

1070 Experiment 3 replicated the key finding from Experiment 2: testing memory immediately

1071 after training significantly boosted retention of new word meanings learned incidentally through

1072 story reading as measured at the delayed tests 24 hours later. Importantly, Experiment 3 showed

1073 that testing memory immediately after training using *either* cued recall or multiple-choice alone

1074 was sufficient to result in a significant testing effect. This is consistent with studies that have

1075 found a testing effect arising from an immediate cued recall test (Karpicke & Smith, 2012) or an

1076 immediate test using multiple-choice questions (Roediger & Marsh, 2005).

1077 The results provide no strong evidence for a particular benefit for either of these two

1078 testing methods: the retention benefits following the immediate multiple-choice test were non-

1079 significantly larger than the immediate cued recall test. The simple effects showed that the

1080 immediate multiple-choice test significantly boosted performance on both of the delayed tests,

1081 while the immediate cued recall test only significantly enhanced performance on the delayed

1082 cued recall test but not on the delayed multiple-choice test (the effect was marginally significant

1083 at the uncorrected level).

1084   Finally, the immediate multiple-choice test enhanced delayed cued recall of the new word

1085   meanings. This cross-task transfer effect suggests that the observed benefit is not simply due to

1086   practising the same test previously. This demonstrates that knowledge retained from prior testing

1087   can be flexibly applied to new contexts of retrieval, in line with the findings of Rohrer et al.

1088   (2010) and others.

1089

1090   **General discussion**

1091   The present experiments investigated participants' ability to learn new word meanings

1092   from naturalistic story contexts. Specifically we (i) compared performance in this relatively

1093   incidental learning condition to a more conventional explicit learning paradigm and (ii)

1094   examined the role of testing memory after training in enhancing future retention of new word

1095   meanings.

1096

1097   **Incidental versus intentional learning**

1098   In Experiment 1 and Experiment 2 participants learned new meanings for familiar words

1099   better under intentional learning conditions than incidentally through reading stories. These

1100   findings are in line with those of studies that have compared incidental and intentional learning

1101   in studies of L2 vocabulary learning (Hulstijn, 1992; Peters et al., 2009) and L1 vocabulary

1102   learning with adolescents (Konopak et al., 1987).

1103   The intentional and incidental learning paradigms differed in several important ways,

1104   making it difficult to draw firm conclusions for the specific underlying cause of this difference. It

1105   may be driven by additional attentional focus on word meanings in the intentional learning

1106   condition, while in the incidental learning condition participants' attention was directed towards

1107   other aspects of the rich narrative context of the stories. The two learning paradigms also differed

1108   in the spacing of the words, which were systematically spaced throughout the intentional

1109   learning task. In the stories, on the other hand, the new word meanings appeared at naturally-

1110   occurring intervals such that some exposures occurred relatively close together. Spacing stimuli

1111   apart has been widely shown to aid learning (for review see: Dempster, 1996).

1112   Additionally, there is the possibility of an internal testing effect within the intentional

1113   learning task. The multiple-choice meaning-to-word matching portion of the intentional learning

1114   task was similar to the multiple-choice meaning-to-word matching task used in the testing phase,

1115   which produced a testing effect on its own in Experiment 3. The multiple-choice training task

1116   also included simple feedback on performance ("correct" or "incorrect"), and feedback can

1117   enhance the benefit of tests for future retention. The trend in the data of Experiment 2 towards a

1118   larger testing effect for the incidental learning condition could possibly be because the

1119   intentional learning condition already involved a limited testing effect.

1120   As well as the overall differences in performance between learning conditions

1121   immediately after training, there were also differences in longer-term retention. After 24 hours,

1122   participants in Experiment 1 (and non-significantly in Experiment 2) had forgotten some of the

1123   new word meanings learned under intentional conditions, but there was very little forgetting of

1124   items learned incidentally across both experiments. In Experiment 2 the trend for reduced

1125   forgetting of items learned under incidental conditions was present regardless of an immediate

1126   test. The significant interaction between day and learning condition in Experiment 1 provides

1127   evidence for a difference in the amount of forgetting: word meanings learned incidentally,

1128   although harder to learn initially, may be forgotten less quickly than those learned under

1129   intentional conditions. This is possibly due to the more semantically rich context of the stories

1130   providing participants with additional and more varied cues, which are advantageous for later

1131   retrieval of the new word meanings, or due to higher engagement as the stories are more

1132   enjoyable. An alternative possibility is that this difference is a function of the initial learning

1133   level as performance in the intentional learning condition was higher and so had further to fall.

1134   Another possibility is that memory consolidation during sleep may have played a role in

1135   preferentially strengthening the weaker memory traces of items acquired through incidental

1136   conditions (Drosopoulos et al., 2007). Nevertheless, while intentional learning conditions were

1137   better for more efficient acquisition, incidental learning may lead to less forgetting of word

1138   meanings over time. However, replication of the latter finding is warranted in future research in a

1139   design that matches on initial performance level between the different learning conditions. This

1140   may also help to determine whether the observed changes in performance over time reflect real

1141   differences in forgetting or differences in the two conditions in terms of different sensitivity to

1142   recall thresholds following the immediate test (see Kornell et al., 2011 for a discussion). Future

1143   replication of this finding is also important given the non-significant interaction between learning

1144   condition and immediate versus delayed first test in Experiment 2.

1145

1146   **The testing effect**

1147   A large overall testing effect was found in both Experiment 2 and Experiment 3: retrieval

1148   practice following initial exposure boosted retention of new meanings for familiar words. Cued

1149   recall of the new word meanings was boosted by 28.8% for the incidental learning condition and

1150   10.6% for the intentional learning condition in Experiment 2. This effect therefore likely

1151   explains the high levels of cued recall and multiple-choice accuracy in Experiment 1 after one

1152   day, and in Hulme et al.'s (2019) study after seven days. This finding adds to the growing

1153   literature highlighting the role of testing in aiding vocabulary learning. A possible alternative

1154   explanation for this finding is that testing may bias participants towards selectively remembering

1155   tested items, so the effect may be driven by the cost to the untested items. Relatedly, research has

1156   shown that the retention of selected memories can be modulated after learning using simple

1157   verbal instructions on their future importance (van Dongen et al., 2012). However, several

1158   studies in the literature using a between-subjects design (e.g., Experiment 2 of Roediger &

1159   Karpicke, 2006a) have demonstrated that the practice of testing enhances learning, rather than

1160   just biasing which items are remembered.

1161   The testing effect in Experiment 2 was elicited with both a cued recall and multiple-

1162   choice test immediately after training. However, the findings of Experiment 3 illustrated that

1163   either a cued recall test or a multiple-choice test alone was sufficient to produce a significant

1164   testing effect following incidental learning through story reading. There was no clear evidence

1165   that either of the two testing methods was superior to the other for enhancing memory retention,

1166   although the retention benefits for the immediate multiple-choice test were non-significantly

1167   larger than for the cued recall test. Previous research has shown cued recall (short answer

1168   questions) to be more helpful for memory retention in some contexts (Duchastel, 1981;

1169   McDaniel, Anderson, et al., 2007; McDaniel, Roediger, et al., 2007). However, recent research

1170   has suggested that the relative benefit of the testing method may depend on the type of

1171   information being learned. Nakata (2016) found that recall was most helpful for acquiring novel

1172   words' orthography (spelling), whereas recognition was more beneficial otherwise.

1173   Despite the growing body of research on the benefits of retrieval practice for retention,

1174   the neurocognitive mechanisms underlying the testing effect remain somewhat unclear (Antony

1175   et al., 2017). Influential models of word learning have not yet provided an account of the testing

1176 effect. The Complementary Learning Systems (CLS) model of word learning, for example,

1177 describes how word forms are initially encoded into episodic memory in the hippocampus, and

1178 are integrated into semantic memory in the neocortex following a period of offline consolidation,

1179 such as during sleep (Davis & Gaskell, 2009). However, this model does not account for the

1180 effect of conscious retrieval on memory for new words and their meanings. Antony et al. (2017)

1181 recently suggested that a similar mechanism may underlie both offline consolidation and the

1182 testing effect, which may provide a fast track to consolidation. They argue that retrieval practice

1183 brings about the formation of flexible hippocampal-neocortical representations through the

1184 online reactivation of related knowledge (Antony et al., 2017). The testing effect is therefore

1185 important to consider in conjunction with offline consolidation processes to garner a full picture

1186 of how novel word meanings are remembered. Future research should explore whether the

1187 testing effect involves a similar mechanism as unconscious offline learning processes, thus

1188 providing a fast track to consolidation.

1189 The findings of the present experiments have important methodological implications for

1190 studies of word learning. The enhancing effects of retrieval practice on memory are clearly

1191 shown here, and in other previous research. Studies considering the impact of factors such as the

1192 role of sleep for consolidation should (and do) consider this important aspect. For example,

1193 Henderson et al. (2015) compared adults' and children's explicit memory of new words using

1194 cued recall and recognition tests administered both immediately and 24 hours later. They note

1195 that for both adults and children "explicit phonological memory was enhanced after off-line

1196 consolidation" p.413 (Henderson et al., 2015). However, this finding could also, at least in part,

1197 be attributable to a testing effect, as the 24-hour tests repeated the 0-hour tests that participants

1198 completed immediately after training. Nevertheless, other studies in this field have isolated

1199 effects of sleep from effects of testing using alternative designs, including train twice, test once

1200 (e.g., Weighall et al., 2017), and AM-PM designs to compare 12-hour periods associated with

1201 wake and sleep (e.g., Henderson et al., 2012; James et al., 2020). The testing effect is therefore

1202 an important consideration for studies of the cognitive mechanisms underlying vocabulary

1203 learning and retention that include repeated testing of trained words. Studies of sleep and

1204 vocabulary learning would benefit from using designs that avoid having multiple test sessions.

1205 For example, training different items at different times and testing all items in one final session

1206 (e.g., Experiment 1 of Tamminen & Gaskell, 2013), or using a between-groups design to

1207 compare between two different times of test (e.g., Experiment 1 of Tamminen et al., 2015)

1208 avoids contaminating results of potential consolidation with those of a testing effect.

1209   The testing effect has previously been shown to generalise to educational settings. For

1210 example, Roediger et al. (2011) found that repeated testing of real course content with multiple-

1211 choice (and some short-answer tests) successfully boosted middle school students' grades on

1212 their social studies course. A similar study with a middle school science class (McDaniel et al.,

1213 2011) found that multiple-choice tests gave large gains (13-25%) in learning and retention,

1214 assessed by end-of-unit exams, especially when tests were taken closer to exam time. Similar

1215 benefits of retrieval practice have been seen for college students (McDaniel, Anderson, et al.,

1216 2007; McDaniel, Roediger, et al., 2007), and Larsen et al. (2008) have advocated the use of test-

1217 enhanced learning in medical education. Our findings in Experiment 2 and Experiment 3 have

1218 important practical implications for vocabulary learning. Students learning vocabulary

1219 incidentally from reading storybooks or textbooks can benefit from being tested following initial

1220 encounters with new word meanings. Testing appears to be effective using either cued recall or

1221 multiple-choice methods, so incorporating it as part of a strategy for efficient vocabulary

1222   learning could be easy to implement. Tests are often considered solely as tools to assess learning,

1223   however they also provide an important opportunity for additional learning and reinforcement of

1224   knowledge.

1225

1226   **Conclusions**

1227   The first two experiments confirmed that new word meanings are learned more

1228   efficiently under intentional learning conditions than incidentally through story reading.

1229   However, there was also some evidence of less forgetting of items learned through stories,

1230   suggesting that word meanings learned in a more semantically rich context could be retained

1231   better. The second two experiments showed that testing memory aids retention of new word

1232   meanings acquired under either incidental or intentional learning conditions. Both cued recall

1233   and recognition (multiple-choice) tests enhanced retention, but multiple-choice tests gave non-

1234   significantly better performance, even with no feedback. Furthermore the testing effect

1235   transferred across test tasks: immediate multiple-choice meaning-to-word matching improved

1236   accuracy on the delayed cued recall test, so the effect is not restricted to benefitting the same test.

1237   The present study has demonstrated that testing memory following initial exposure is a

1238   powerful way to improve learning and long-term retention of vocabulary knowledge.

1239   Importantly, we found that retrieval practice benefitted vocabulary retention from different

1240   learning conditions and using different methods of immediate test. The vast majority of new

1241   vocabulary is learned through reading from mid-childhood onwards. Test-enhanced learning

1242   could therefore be particularly useful if implemented during vocabulary development to boost

1243   children's vocabulary gains from story reading.

1244

Acknowledgements

1246   The authors would like to thank Helen Moss and Johan Heemskerk for authoring the

1247 stories used as stimulus materials, and Rachel Jose for developing the comprehension questions

1248 used with the stories and assisting with coding the data. The authors would also like to thank Dr

1249 Eva Poort for providing feedback on an earlier version of this article.

References

1251 Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to

1252   memory consolidation. *Trends in Cognitive Sciences*, *21*(8), 573–576.

1253   https://doi.org/10.1016/j.tics.2017.05.001

1254 Armstrong, R., Arnott, W., Copland, D. A., McMahon, K., Khan, A., Najman, J. M., & Scott, J.

1255   G. (2017). Change in receptive vocabulary from childhood to adulthood: associated mental

1256   health, education and employment outcomes. *International Journal of Language and*

1257   *Communication Disorders*, *52*(5), 561–572. https://doi.org/10.1111/1460-6984.12301

1258 Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for

1259   confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3),

1260   255–278. https://doi.org/10.1016/j.jml.2012.11.001

1261 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models

1262   using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

1263   https://doi.org/10.18637/jss.v067.i01

1264 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2016). *lme4: Linear mixed-effects models*

1265   *using "eigen" and s4*. [Software Manual]. https://cran.r-

1266   project.org/web/packages/lme4/lme4.pdf

1267 Batterink, L., & Neville, H. (2011). Implicit and explicit mechanisms of word learning in a

1268 narrative context: An event-related potential study. *Journal of Cognitive Neuroscience,*

1269 *23*(11), 3181–3196. https://doi.org/10.1162/jocn_a_00013

1270 Betts, H. N., Gilbert, R. A., Cai, Z. G., Okedara, Z. B., & Rodd, J. M. (2018). Retuning of

1271 lexical-semantic representations: Repetition and spacing effects in word-meaning priming.

1272 *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(7), 1130–1150.

1273 https://doi.org/10.1037/xlm0000507

1274 Blything, L. P., Hardie, A., & Cain, K. (2020). Question Asking During Reading Comprehension

1275 Instruction: A Corpus Study of How Question Type Influences the Linguistic Complexity of

1276 Primary School Students' Responses. *Reading Research Quarterly*, *55*(3), 443–472.

1277 https://doi.org/10.1002/rrq.279

1278 Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E.,

1279 & Rosen, B. R. (1996). Detection of cortical activation during averaged single trials of a

1280 cognitive task using functional magnetic resonance imaging. *Proceedings of the National*

1281 *Academy of Sciences*, *93*(25), 14878–14883. https://doi.org/10.1073/pnas.93.25.14878

1282 Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated

1283 studying. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(5),

1284 1118–1133. https://doi.org/10.1037/a0019902

1285 Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L. (2006). When additional multiple-

1286 choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*(7), 941–

1287 956. https://doi.org/10.1002/acp.1239

1288 Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more

1289 important for some aspects of comprehension? *L'Annee Psychologique*, *114*(4), 647–662.

1290 https://doi.org/10.4074/S0003503314004035

1291   Carpenter, S. K. (2009). Cue Strength as a Moderator of the Testing Effect: The Benefits of

1292   Elaborative Retrieval. *Journal of Experimental Psychology: Learning Memory and*

1293   *Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

1294   Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to

1295   Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–

1296   45. https://doi.org/10.20982/tqmp.01.1.p042

1297   Damer, E., & Bradley, P. (2014). *Prolific Academic [Computer Software]*.

1298   https://www.prolific.ac/

1299   Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning:

1300   Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of*

1301   *London. Series B, Biological Sciences*, *364*, 3773–3800.

1302   https://doi.org/10.1098/rstb.2009.0111

1303   Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E.

1304   L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (pp. 31–

1305   344). Academic Press.

1306   Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's Function in the Spontaneous

1307   Recovery and Consolidation of Memories. *Journal of Experimental Psychology: General*,

1308   *136*(2), 169–183. https://doi.org/10.1037/0096-3445.136.2.169

1309   Duchastel, P. C. (1981). Retention of prose following testing with different types of tests.

1310   *Contemporary Educational Psychology*, *6*(3), 217–226. https://doi.org/10.1016/0361-

1311   476X(81)90002-3

1312   Eligio, R. B., & Kaschak, M. P. (2021). Gaming experience affects the interpretation of

1313   ambiguous words. *PLoS ONE*, *15*(12 December), e0243512.

1314   https://doi.org/10.1371/journal.pone.0243512

1315   Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and

1316   combining retrieval practice and the keyword mnemonic for foreign vocabulary learning.

1317   *Applied Cognitive Psychology*, *21*(4), 499–526. https://doi.org/10.1002/acp.1287

1318   Gaskell, M. G., Cairney, S. A., & Rodd, J. M. (2019). Contextual priming of word meanings is

1319   stabilized over sleep. *Cognition*, *182*, 109–126.

1320   https://doi.org/10.1016/j.cognition.2018.09.007

1321   Gilbert, R. A., Davis, M. H., Gareth Gaskell, M., & Rodd, J. M. (2018). Listeners and readers

1322   generalize their experience with word meanings across modalities. *Journal of Experimental*

1323   *Psychology: Learning, Memory, and Cognition*, *44*(10), 1533–1561.

1324   https://doi.org/10.1037/xlm0000532

1325   Gilbert, R. A., Davis, M. H., Gaskell, M. G., & Rodd, J. M. (2021). The relationship between

1326   sentence comprehension and lexical-semantic retuning. *Journal of Memory and Language*,

1327   *116*, 104188. https://doi.org/10.1016/j.jml.2020.104188

1328   Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A., & Yoon,

1329   H.-J. (2017). Incidental vocabulary learning in a natural reading context: An eye-tracking

1330   study. *Bilingualism: Language and Cognition*, *21*(3), 563–584.

1331   https://doi.org/10.1017/S1366728917000219

1332   Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of

1333   retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology*,

1334   *28*(1), 135–142. https://doi.org/10.1002/acp.2956

1335   Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A.

1336   (2014). The benefit of retrieval practice over elaborative restudy in primary school

1337   vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*(3), 177–182.

1338   https://doi.org/10.1016/j.jarmac.2014.05.003

1339 Henderson, L. M., Devine, K., Weighall, A., & Gaskell, G. (2015). When the daffodat flew to

1340   the intergalactic zoo: Off-line consolidation is critical for word learning from stories.

1341   *Developmental Psychology*, *51*(3), 406–417. https://doi.org/10.1037/a0038786

1342 Henderson, L. M., Weighall, A. R., Brown, H., & Gaskell, M. G. (2012). Consolidation of

1343   vocabulary is associated with sleep in children. *Developmental Science*, *15*(5), 674–687.

1344   https://doi.org/10.1111/j.1467-7687.2012.01172.x

1345 Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term

1346   recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562–567.

1347   https://doi.org/10.1016/S0022-5371(71)80029-4

1348 Hulme, R. C., Barsky, D., & Rodd, J. M. (2019). Incidental Learning and Long-Term Retention

1349   of New Word Meanings From Stories: The Effect of Number of Exposures. *Language*

1350   *Learning*, *69*(1), 18–43. https://doi.org/10.1111/lang.12313

1351 Hulme, R. C., & Rodd, J. M. (2016). *The testing effect in long-term retention of novel meanings*

1352   *for known words learned through incidental and intentional means*. https://osf.io/e5zmk/

1353 Hulme, R. C., & Rodd, J. M. (2017). *The impact of test method on the testing effect in long-term*

1354   *retention of new word meanings learned incidentally from stories*. https://osf.io/c59tz/

1355 Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental

1356   vocabulary learning. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied*

1357   *linguistics* (pp. 113–125). Macmillan. https://doi.org/10.1007/978-1-349-12396-4_11

1358 Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.),

1359   *The handbook of second language acquisition* (pp. 349–381). Blackwell.

1360   https://doi.org/10.1002/9780470756492.ch12

1361   James, E., Gaskell, M. G., & Henderson, L. M. (2020). Sleep-dependent consolidation in

1362   children with comprehension and vocabulary weaknesses: it'll be alright on the night?

1363   *Journal of Child Psychology and Psychiatry*, *61*(10), 1104–1115.

1364   https://doi.org/10.1111/jcpp.13253

1365   Kang, S. H. K., Mcdermott, K. B., Roediger, H. L., & Kang, S. (2007). Test format and

1366   corrective feedback modify the effect of testing on long-term retention. *European Journal*

1367   *of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

1368   Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term

1369   retention, but equally spaced retrieval enhances long-term retention. *Journal of*

1370   *Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704–719.

1371   https://doi.org/10.1037/0278-7393.33.4.704

1372   Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrival for learning.

1373   *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

1374   Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and

1375   elaborative encoding. *Journal of Memory and Language*, *67*(1), 17–29.

1376   https://doi.org/10.1016/J.JML.2012.02.004

1377   Konopak, B., Sheard, C., Longman, D., Lyman, B., Slaton, E., Atkinson, R., & Thames, D.

1378   (1987). Incidental versus intentional word learning from context. *Reading Psychology: An*

1379   *International Quarterly*, *8*(1), 7–21. https://doi.org/10.1080/0270271870080103

1380   Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A

1381   distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97.

1382   https://doi.org/10.1016/j.jml.2011.04.002

1383   Larsen, D. P., Butler, A. C., & Roediger, H. L. (2008). Test-enhanced learning in medical

1384   education. *Medical Education*, *42*(10), 959–966. https://doi.org/10.1111/j.1365-

1385   2923.2008.03124.x

1386   Lehmann, M. (2007). Is intentional or incidental vocabulary learning more effective? *The*

1387   *International Journal of Foreign Language Teaching*, *3*(1), 23–28.

1388   Maciejewski, G., & Klepousniotou, E. (2020). Disambiguating the ambiguity disadvantage

1389   effect: Behavioral and electrophysiological evidence for semantic competition. *Journal of*

1390   *Experimental Psychology: Learning Memory and Cognition*, *46*(9), 1682–1700.

1391   https://doi.org/10.1037/xlm0000842

1392   Maciejewski, G., Rodd, J. M., Mon-Williams, M., & Klepousniotou, E. (2020). The cost of

1393   learning new meanings for familiar words. *Language, Cognition and Neuroscience*, *35*(2),

1394   188–210. https://doi.org/10.1080/23273798.2019.1642500

1395   Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences

1396   of multiple-choice testing. *Psychonomic Bulletin and Review*, *14*(2), 194–199.

1397   https://link.springer.com/content/pdf/10.3758/BF03194051.pdf

1398   Mason, B., & Krashen, S. (2004). Is form-focused vocabulary instruction worth while? *RELC*

1399   *Journal*, *35*(2), 179–185. https://doi.org/10.1177/003368820403500206

1400   McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011).

1401   Test-Enhanced Learning in a Middle School Science Classroom: The Effects of Quiz

1402   Frequency and Placement. *Journal of Educational Psychology*, *103*(2), 399–414.

1403   https://doi.org/10.1037/a0021782

1404   McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing

1405   effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513.

1406   https://doi.org/10.1080/09541440701326154

1407 McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced

1408   learning from the laboratory to the classroom. *Psychonomic Bulletin and Review*, *14*(2),

1409   200–206. https://link.springer.com/content/pdf/10.3758/BF03194052.pdf

1410 Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning Word Meanings From Context

1411   During Normal Reading. *American Educational Research Journal Summer*, *24*(2), 237–

1412   270. http://journals.sagepub.com/doi/pdf/10.3102/00028312024002237

1413 Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading*

1414   *Research Quarterly*, *20*(2), 233–253. http://www.jstor.org/stable/747758

1415 Nakata, T. (2016). Effects of retrieval formats on second language vocabulary learning.

1416   *International Review of Applied Linguistics in Language Teaching*, *54*(3), 257–289.

1417   https://doi.org/10.1515/iral-2015-0022

1418 Nation, K. (2017). Nurturing a lexical legacy: Reading experience is critical for the development

1419   of word reading skill. *Npj Science of Learning*, *2*(1), 3. https://doi.org/10.1038/s41539-017-

1420   0004-7

1421 Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate

1422   learning of words? *Journal of Experimental Psychology: Learning Memory and Cognition*,

1423   *31*(1), 3–8. https://doi.org/10.1037/0278-7393.31.1.3

1424 Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading.

1425   *Studies in Second Language Acquisition*, *38*(1), 97–130.

1426   https://doi.org/10.1017/S0272263115000224

1427 Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary

1428   through reading: The effect of three enhancement techniques compared. *Language*

1429   *Learning*, *59*(1), 113–151. https://doi.org/10.1111/j.1467-9922.2009.00502.x

1430   Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater

1431   difficulty correctly recalling information lead to higher levels of memory? *Journal of*

1432   *Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

1433   Qualtrics. (2015). *Qualtrics Survey Software*. Qualtrics. https://www.qualtrics.com

1434   R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation

1435   for Statistical Computing, Vienna, Austria. https://www.r-project.org/

1436   Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and*

1437   *vocabulary scales. Section 5: The Mill Hill vocabulary scale*. Harcourt Assessment.

1438   Rodd, J. M. (2018). Lexical Ambiguity. In S.-A. Rueschemeyer & M. G. Gaskell (Eds.), *Oxford*

1439   *Handbook of Psycholinguistics* (2nd ed., pp. 96–117). Oxford University Press.

1440   Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-

1441   Meaning Access. *Perspectives on Psychological Science*, *15*(2), 411–427.

1442   https://doi.org/10.1177/1745691619885860

1443   Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012).

1444   Learning new meanings for old words: Effects of semantic relatedness. *Memory &*

1445   *Cognition*, *40*(7), 1095–1108. https://doi.org/10.3758/s13421-012-0209-1

1446   Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact

1447   of recent and long-term experience on access to word meanings: Evidence from large-scale

1448   internet-based experiments. *Journal of Memory and Language*, *87*, 16–37.

1449   https://doi.org/10.1016/J.JML.2015.10.006

1450   Rodd, J. M., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity:

1451   Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266.

1452   https://doi.org/10.1006/jmla.2001.2810

1453   Rodd, J. M., Lopez Cutrin, B., Kirsch, H., Millar, A., & Davis, M. H. (2013). Long-term priming

1454   of the meanings of ambiguous words. *Journal of Memory and Language*, *68*(2), 180–198.

1455   https://doi.org/10.1016/j.jml.2012.08.002

1456   Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-Enhanced

1457   Learning in the Classroom: Long-Term Improvements From Quizzing. *Journal of*

1458   *Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

1459   Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term

1460   retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

1461   https://doi.org/10.1016/j.tics.2010.09.003

1462   Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests

1463   improves long-term retention. *Psychological Science*, *17*(3), 249–255.

1464   https://doi.org/10.1111/j.1467-9280.2006.01693.x

1465   Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and

1466   implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–

1467   210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

1468   Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-

1469   choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,

1470   *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

1471   Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of*

1472   *Experimental Psychology: Learning Memory and Cognition*, *36*(1), 233–239.

1473   https://doi.org/10.1037/a0017678

1474   Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review

1475   of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463.

1476   https://doi.org/10.1037/a0037559

1477 Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*,

1478   *6*(2), 72–78. https://doi.org/10.1016/0346-251X(78)90027-1

1479 Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge

1480   in language learning. *Cognitive Psychology*, *79*, 1–39.

1481   https://doi.org/10.1016/j.cogpsych.2015.03.003

1482 Tamminen, J., & Gaskell, M. G. (2013). Novel word integration in the mental lexicon: evidence

1483   from unmasked and masked semantic priming. *Quarterly Journal of Experimental*

1484   *Psychology (2006)*, *66*(5), 1001–1025. https://doi.org/10.1080/17470218.2012.724694

1485 Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions

1486   and answers. *Experimental Psychology*, *56*, 252–257. https://doi.org/10.1027/1618-

1487   3169.56.4.252

1488 Tran, R., Rohrer, D., & Pashler, H. (2014). Retrieval practice: The lack of transfer to deductive

1489   inferences. *Psychonomic Bulletin & Review*, *22*(1), 135–140.

1490   https://doi.org/10.3758/s13423-014-0646-x

1491 van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013).

1492   Neural correlates of testing effects in vocabulary learning. *NeuroImage*, *78*, 94–102.

1493   https://doi.org/10.1016/j.neuroimage.2013.03.071

1494 van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual richness

1495   and word learning: Context enhances comprehension but retrieval enhances retention.

1496   *Language Learning*, *68*(2), 546–585. https://doi.org/10.1111/lang.12285

1497 van Dongen, E. V., Thielen, J.-W., Takashima, A., Barth, M., & Fernández, G. (2012). Sleep

…

…

…

1498   Supports Selective Retention of Associative Memories Based on Relevance for Future

1499   Utilization. *PLoS ONE*, *7*(8), e43426. https://doi.org/10.1371/journal.pone.0043426

1500   Weighall, A. R., Henderson, L. M., Barr, D. J., Cairney, S. A., & Gaskell, M. G. (2017). Eye-

1501   tracking the time-course of novel word learning and lexical competition in adults and

1502   children. *Brain and Language*, *167*, 13–27. https://doi.org/10.1016/j.bandl.2016.07.010

1503

Figure 1

Experiment 1. Mean percentage of correct responses on the cued recall test for each learning condition, when tested on day one (immediately after learning) and 24 hours later (N = 31).

The LME analyses were carried out on the raw binary accuracy data, however mean

percentage accuracy data are displayed in the graphs for ease of interpretation. Error bars

show standard errors for the means, adjusted for the within-participant design (Cousineau,

2005).
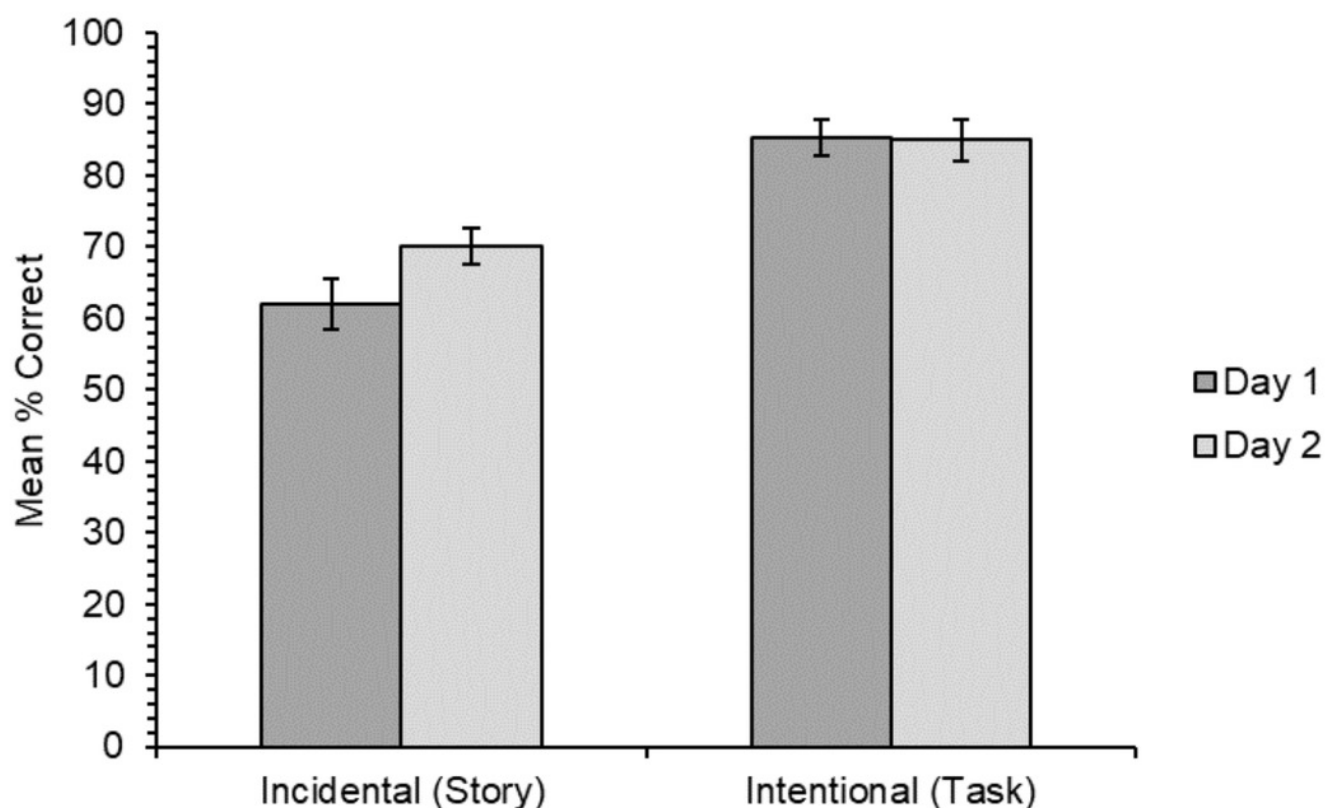
Figure 2

Experiment 1. Mean percentage of correct responses on the multiple-choice test for each learning condition, when tested on day one (immediately after learning) and day two (24 hours later; $N = 31$).

Error bars show standard errors for the means, adjusted for the within-participant factor of learning condition (Cousineau, 2005).
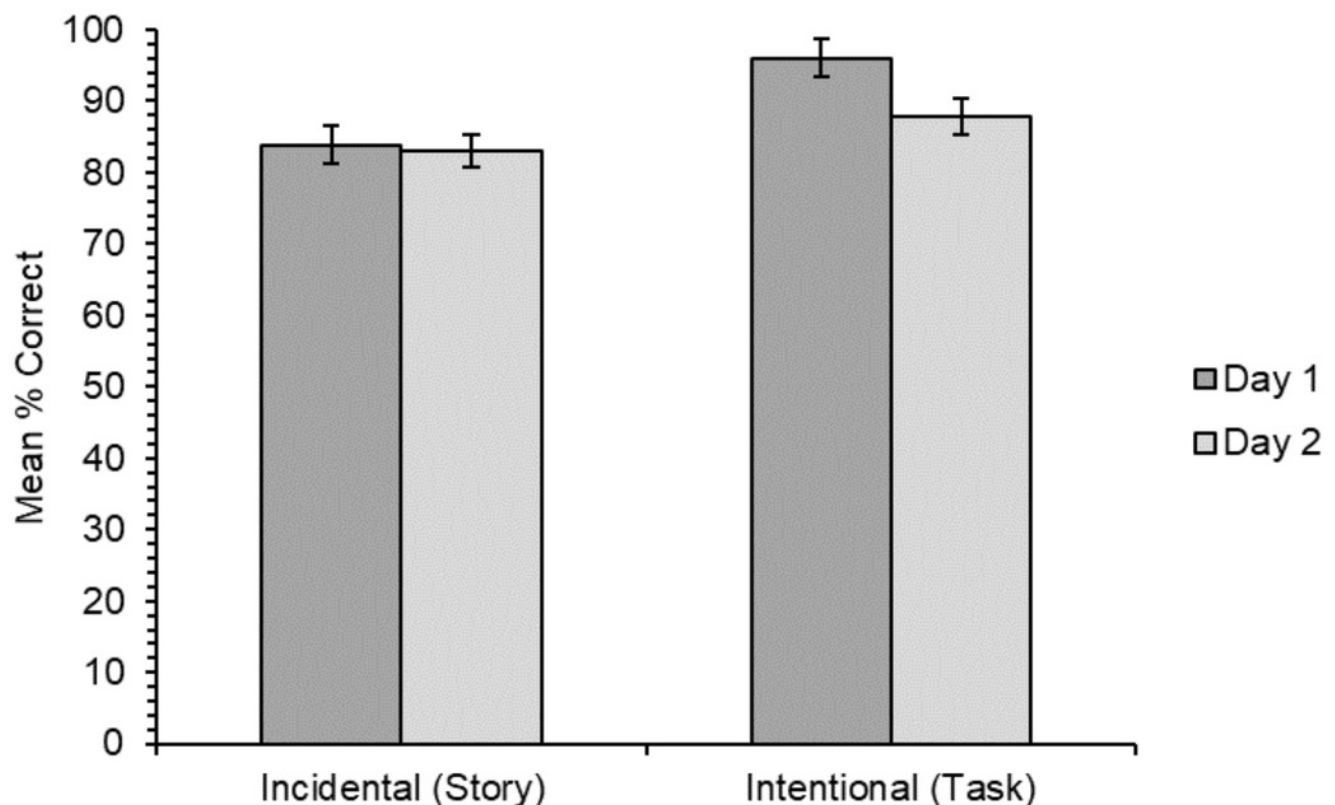
Figure 3

Experiment 2. Mean percentage of correct responses in the cued recall test for each learning condition and for the three different test types in the experiment.

Error bars show standard error of the means adjusted for the within-participant design (Cousineau, 2005).
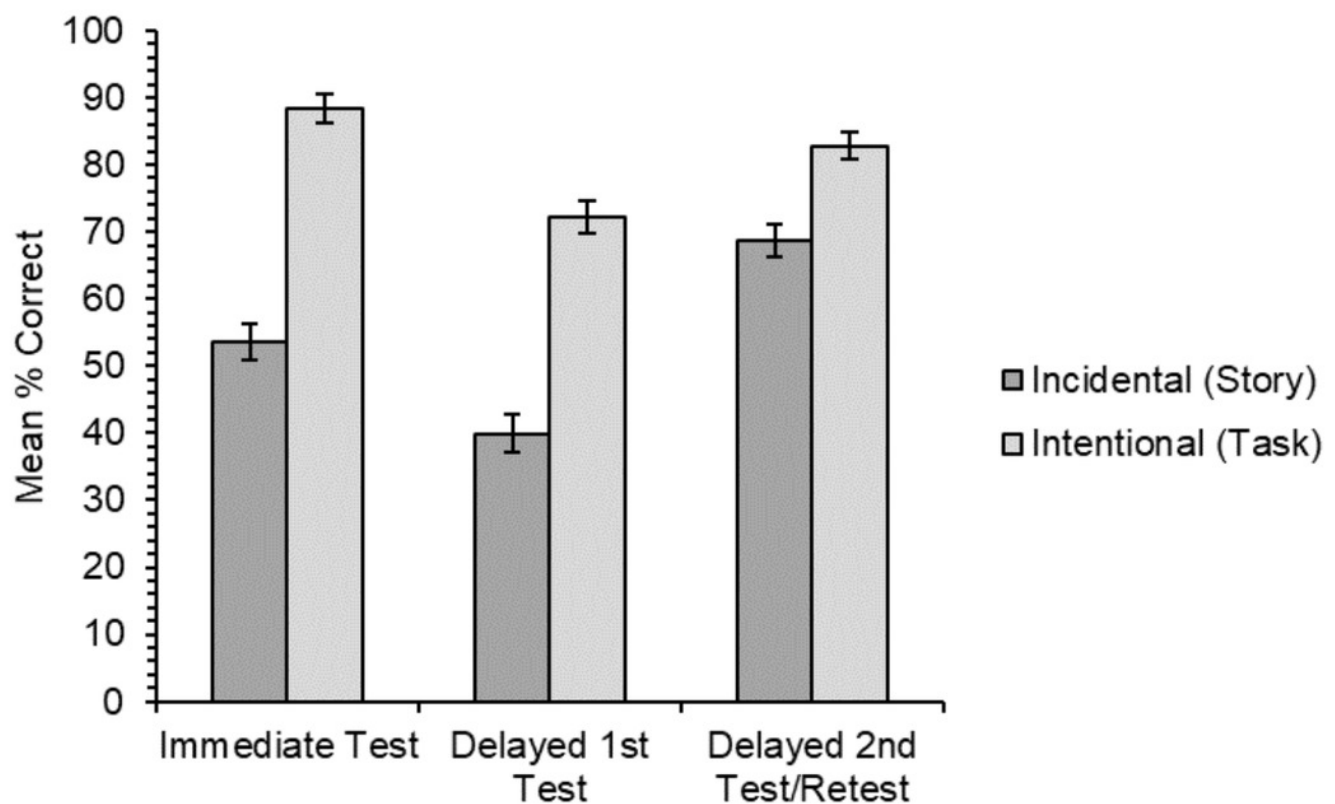
Table 1. Results for the linear mixed effects model analyses for the cued recall test in Experiment 2 showing the main analyses and the different levels of follow-up analyses.

Test results reported are for likelihood ratio tests comparing the full model to models with each fixed factor/interaction of interest removed in turn.

Fixed effect or interaction	χ^2	df	p
1. Main analysis (α = .05)			
Learning condition (Incidental vs. Intentional)	34.83	1	<.001
Test type (Immediate vs. Delayed 1st vs. Delayed 2nd)	25.78	2	<.001
Learning order (First task vs. Second task)	2.50	1	.114
Learning condition x Test type	13.86	2	<.001
Learning condition x Learning order	0.09	1	.760
Test type x Learning order	9.24	2	.010
Learning condition x Test type x Learning order	3.23	2	.199
2. Follow-up pairwise comparisons of the different levels of Test type (α = .017)			
Immediate vs. Delayed 1st tests	19.99	1	<.001
Delayed 1st vs. Delayed 2nd tests	18.83	1	<.001
Immediate vs. Delayed 2nd tests	0.89	1	.345
3. Follow-up pairwise comparisons of 2x2 interactions between pairs of levels of Test type and the two Learning conditions (α = .017)			
Learning condition x Immediate vs. Delayed 2nd tests	16.24	1	<.001
Learning condition x Immediate vs. Delayed 1st tests	2.29	1	.130
Learning condition x Delayed 1st vs. Delayed 2nd tests	2.97	1	.085
4. Simple effects analyses of differences between levels of Test type within the two Learning conditions (α = .008)			
Immediate vs. Delayed 1st tests for Incidental learning	5.88	1	.015

Delayed 1st vs. Delayed 2nd tests for Incidental learning	15.27	1	<.001
Immediate vs. Delayed 2nd tests for Incidental learning	14.59	1	<.001
Immediate vs. Delayed 1st tests for Intentional learning	18.27	1	<.001
Delayed 1st vs. Delayed 2nd tests for Intentional learning	8.39	1	.004
Immediate vs. Delayed 2nd tests for Intentional learning	4.25	1	.039

Exploratory analyses: Pairwise comparisons of 2x2 interactions between pairs of levels of Test type and the two Learning orders (α = .017)

Learning order x Delayed 1st vs. Delayed 2nd tests	7.12	1	.008
Learning order x Immediate vs. Delayed 2nd tests	0.12	1	.729
Learning order x Immediate vs. Delayed 1st tests	5.15	1	.023

Exploratory analyses: Simple effects analyses of differences between levels of Test type within the two Learning orders (α = .025)

First task vs. Second task for Immediate test	2.25	1	.134
First task vs. Second task for Delayed 1st test	2.04	1	.154
First task vs. Second task for Delayed 2nd test	4.91	1	.027

Figure 4

Experiment 2. Mean percentage of correct responses in the multiple-choice test for each learning condition and for the three different test types in the experiment.

Note that the results from the immediate test are not comparable to those from the two delayed test types due to an underlying difference in test difficulty. Error bars show standard error of the means adjusted for the within-participant design (Cousineau, 2005).
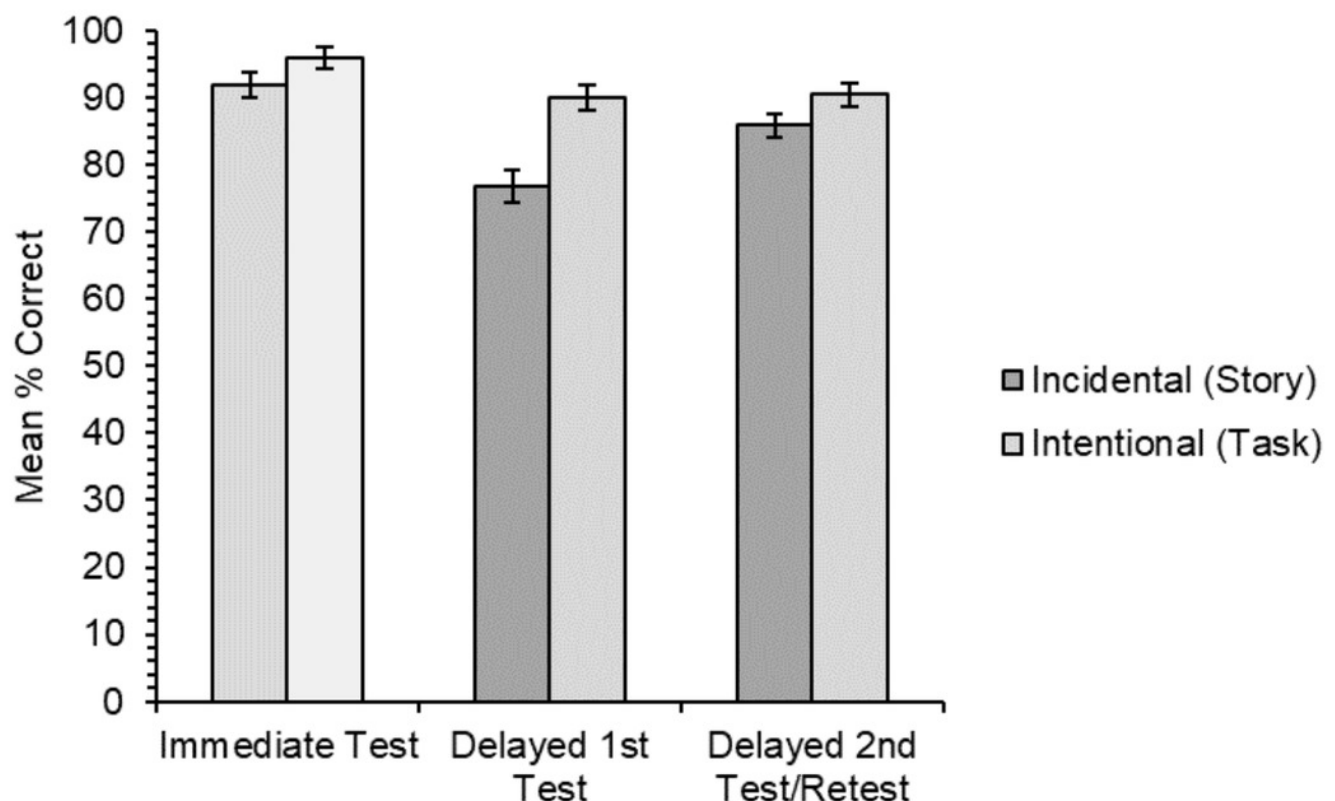
Figure 5

Experiment 3. Mean percentage of correct responses in the cued recall test measured at the delayed test for participants whose immediate test was cued recall, and for those whose immediate test was multiple-choice when items were or were not previously te

Error bars show standard error of the means adjusted for the within-participants factor of whether items were or were not previously tested (Cousineau, 2005).
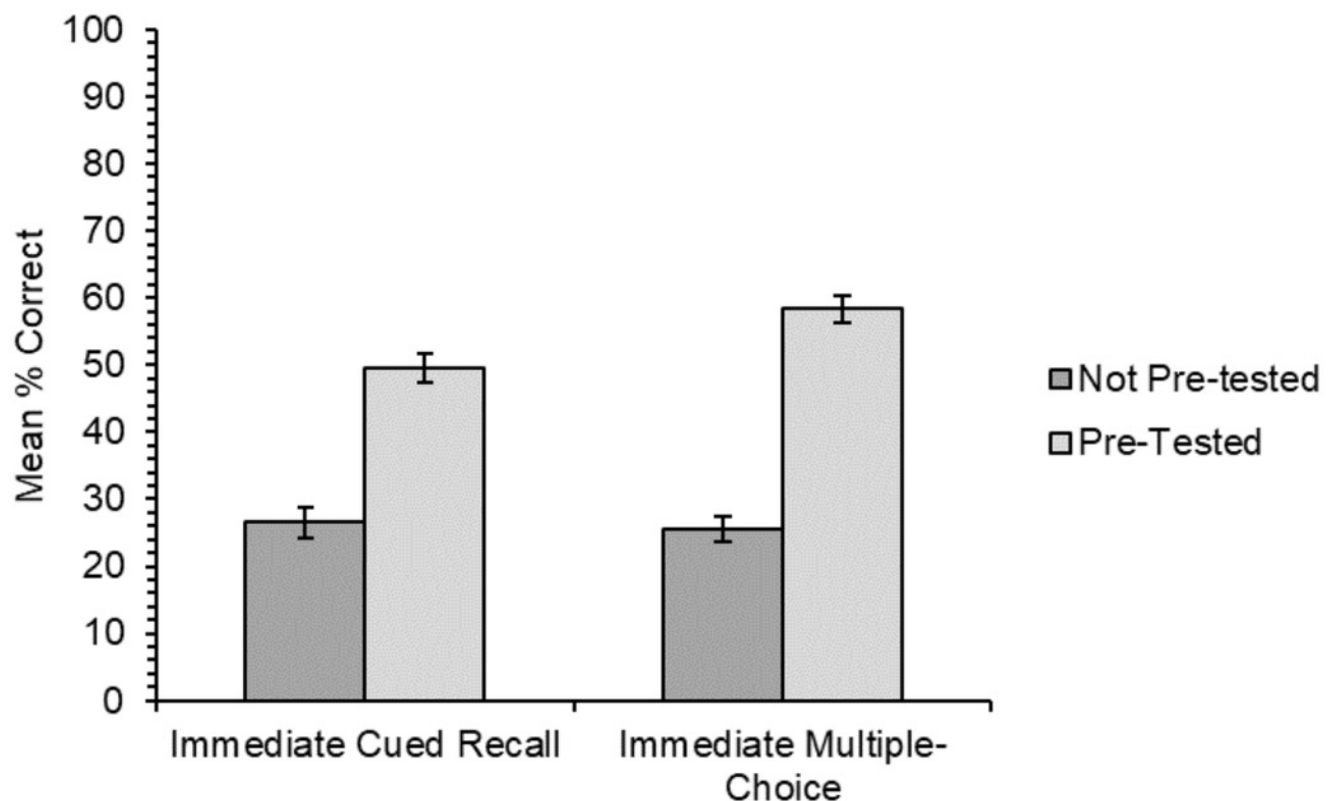
Figure 6

Experiment 3. Mean percentage of correct responses in the multiple-choice test measured at the delayed test for participants whose immediate test was cued recall, and for those whose immediate test was multiple-choice, when items were or were not previous

Error bars show standard error of the means adjusted for the within-participants factor of whether items were or were not previously tested (Cousineau, 2005).