Final Report_Replication of Study 2 by James et al. (2021, JEP:LMC)

Madison Paron (mparon@stanford.edu)

2025-12-13

Table of contents

Project Repo: https://github.com/psych251/james2021.git

Original Paper: https://github.com/psych251/james2021/blob/f321f4e5839f4ffe743184268abcec2df0c944cf/original_paper/james21.pdf

1 Introduction

1.1 Justification for Choice of Experiment

I chose this experiment because it examines how prior lexical knowledge influences vocabulary acquisition under incidental learning conditions, which aligns closely with my research interests in language learning and memory integration. Most models of word learning emphasize explicit instruction, yet the majority of real-world vocabulary acquisition occurs incidentally through narrative and social contexts. I am particularly interested in how prior knowledge interacts with consolidation processes to shape long-term lexical representations—a question that sits at the intersection of language acquisition and memory systems research. Replicating this experiment offers an opportunity to examine whether the memory and language mechanisms observed in controlled learning settings extend to more naturalistic, story-based contexts that mirror how vocabulary is acquired in everyday life.

1.2 Description of Stimuli, Procedures, and Anticipated Challenges

The experiment presents 15 bisyllabic pseudowords embedded within an illustrated, spoken story ("*Trouble at the Intergalactic Zoo*"). Each pseudoword belongs to one of three phonological neighborhood conditions:
- **No neighbors** (e.g., *femod*)
- **One neighbor** (e.g., *tabric* *fabric*)
- **Many neighbors** (e.g., *dester* *duster*, *pester*)

Each pseudoword appears five times across the narrative. The story is paired with 15 corresponding cartoon scenes, each containing multiple pseudoword referents to maintain narrative coherence while preventing explicit word–object pairing. This design ensures *incidental exposure* rather than deliberate memorization.

Participants listen to the 7-minute story while viewing illustrations, then complete three types of memory tests:

1. **Stem completion** – recall of word-forms from initial CV cues

2. **Form recognition** – distinguishing target pseudowords from minimal phonological foils

3. **Form–picture recognition** – mapping pseudowords to their illustrated referents

Each test is administered **immediately after learning** and **the next day** to assess consolidation effects.

1.3 Challenges in Replication

- **Stimulus control:** Ensuring balanced pseudoword properties (phoneme/letter length, bigram probability, neighborhood frequency) and matching the auditory timing of exposures. (Would like to figure out how to check this)
- **Incidental exposure fidelity:** Participants must attend to the story without adopting explicit memorization strategies, especially in adult online samples.

- **Retention and engagement:** Maintaining consistent participation across multiple testing sessions, particularly for the delayed tests.

2 Methods

2.1 Power Analysis

Original effect size, power analysis for samples to achieve 80%, 90%, 95% power to detect that effect size. Considerations of feasibility for selecting planned sample size.

2.2 Planned Sample

- 60 adults
- Age 18-35 years old
- Native monolingual English speakers residing in the US (will most likely change this to US English speakers residing in the US)
- No reported visual, hearing, or literacy difficulties
- Had not participated in experiment S1 (not running S1, so not a problem)
- Have working microphone that was compatible with the experiment platform
- Prolific participants that had participated in at least ten studies with minimum 95% approval rate
- Must not self-report inappropriate strategy (i.e., writing the words down)

- Must complete vocabulary test properly

2.3 Materials

This includes audio and images. I have gathered these materials from the posted materials on OSF.

2.4 Procedure

Three Testing Days:

- **Session 1 / Day 1:** ~20 minutes
- **Session 2 / Day 2:** ~5 minutes
- ***No session 3 for retention and financing constraints***

2.5 Analysis Plan

Analyses will be conducted in R, using lme4 to fit mixed effects models and ggplot2 for figures. A mixed effects binomial regression model will be used to analyze each of the dependent variables, with fixed effects of session, neighborhood condition, vocabulary ability, and all corresponding interactions. Orthogonal contrasts will be used for each of the factorial predictors. For the fixed effect of session, delay1 contrasted responses before and after opportunities for offline consolidation (T1 vs. T2). For the fixed effect of neighbors, neighb1 contrasted words without versus with neighbors (no vs. one & many), and neighb2 contrasted words with one versus many neighbors. I will attempt to figure out how the authors used raw vocabulary scores for analyses, which were scaled and centered before entering into the model. For each analysis, I will computed a random-intercepts model with all fixed effects and interactions. If there was no indication of a three-way interaction in the model (all ps $> .2$), these will be pruned to enable a more parsimonious model with better-specified random effects. I will then incorporate random slopes into the model using a forward best-path approach (Barr et al., 2013), progressively adding slopes into the model and retaining only those random effects justified by the data under a liberal alpha-criterion ($p < .2$).

Clarify key analysis of interest here
I am specifically interested in the form recognition analysis (the test where the participant has to listen to two words, one word that was present in the study and a foil word, and then make a determination of which word they think they heard in the study). I think it will be interesting to see how these results differ after the day delay.

2.6 Differences from Original Study

- I will only focus on experiment 2 that focuses on adults (18-35 years of age). I had trouble with a microphone check in my jspsych script, so I decided to push forward using Gorilla (the platform originally used by the authors) and thought it would be a great way to test exactly what their procedures were.
- My participants will be from the US rather from the UK.
- I also implemented an additional exclusion criteria that Prolific offered to return submissions of participants that completed the task unrealistically fast (although I am not sure how the software made this determination).
- I only conducted two test sessions rather than three. The original study demonstrated that neighborhood effects became more pronounced after longer consolidation intervals, particularly at the one-week delay.

Unfortunately, I had lots of issues with participant retention due to a delay which participants had to return the next day to complete the remainder of the study. I would love to continue running the experiment to get enough power to look into the results further, but with the financial constraints, and only being an auditor in the course, I did not want to spend more funds on this single experiment.

2.7 Methods Addendum (Post Data Collection)

In addition to following the methods outlined in the paper, I also sent a Prolific message once participants completed session 1 to inform them that another Prolific task will be made available to them in 22 hours of completion of session 1. The next day, I sent a reminder email to the participants.

2.7.1 Actual Sample

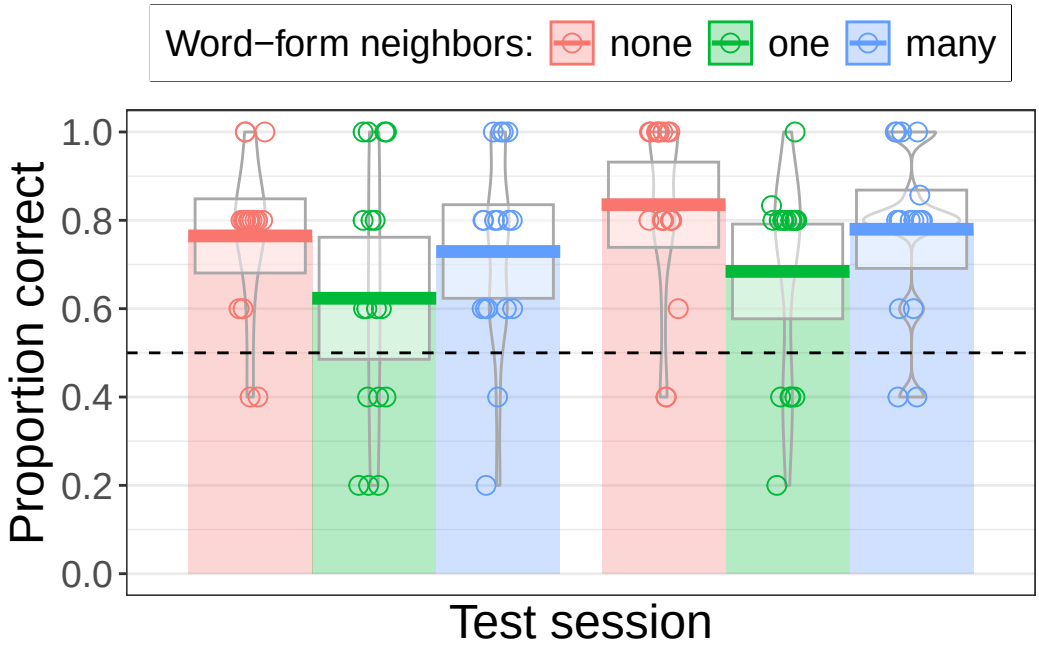
- Number of participants: 17 participants obtained through Prolific
- Age: 18-35 years old
- Residing in the United States
- First Language and Native Language: English (US)
- No reported visual, hearing, or literacy difficulties
- Have working microphone that was compatible with the experiment platform
- Prolific participants that had participated in at least ten studies with minimum 95% approval rate
- Must not self-report inappropriate strategy (i.e., writing the words down)

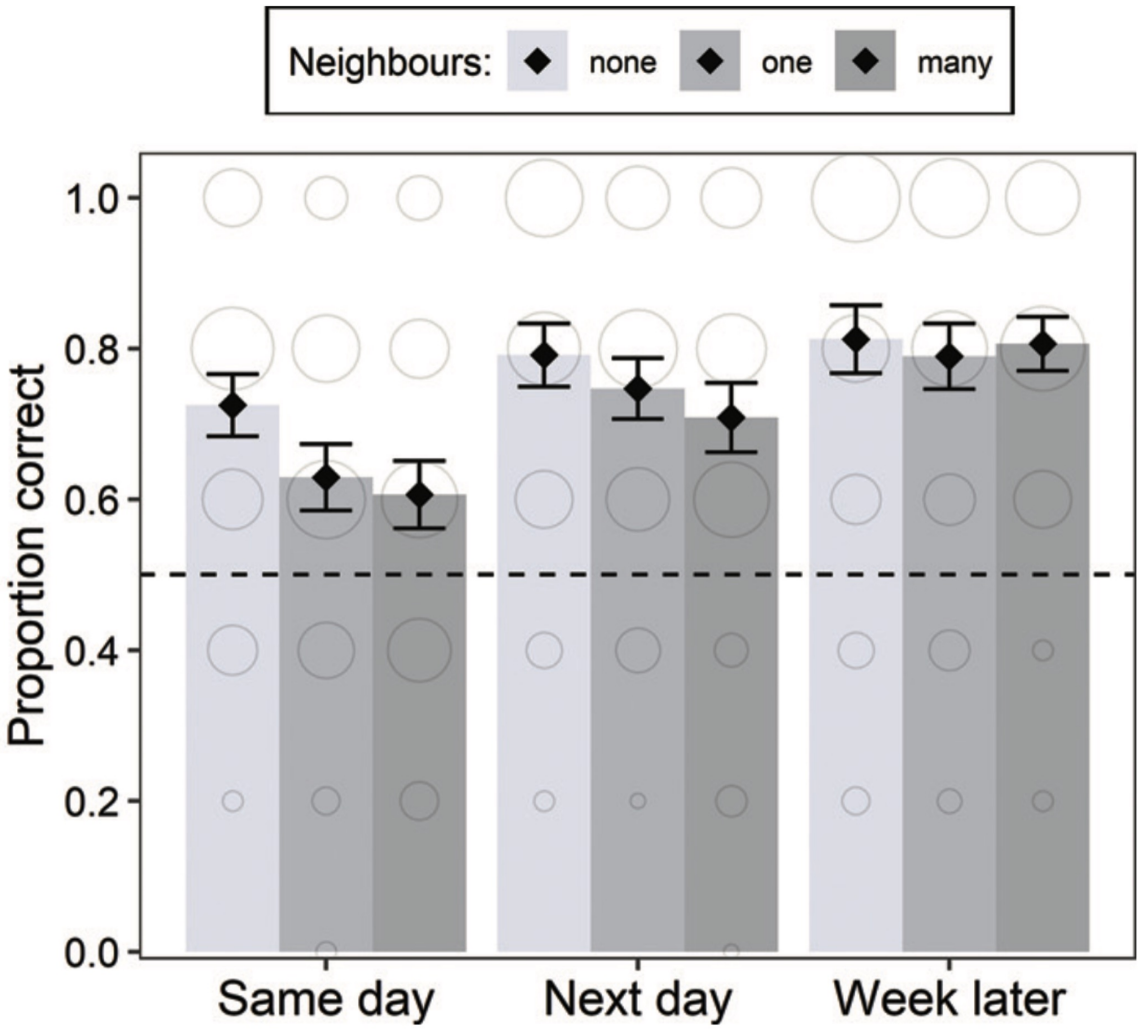
2.7.2 Differences from pre-data collection methods plan

The sample size was substantially smaller than in the original study, limiting statistical power. This is reflected in frequent singular model fits and the inability to justify random slopes, despite following an identical modeling strategy. These issues suggest that the data contained insufficient information to reliably estimate participant-level variability in learning trajectories.

3 Results

```
newgraph <- readRDS("formGraph.rds")
newgraph
```





3.1 Data preparation

Data preparation following the analysis plan.

The current pilot B raw data can be found here.

The current pilot B clean data can be found here.

The drafted analysis script can be found here.

3.2 Confirmatory analysis

3.2.1 Preprocessing scripts

This takes the files from the survey tool and converts them into files that are easier to process.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(purrr)

## ------------------------------------------------------------
## 0. Paths
## ------------------------------------------------------------

base_dir <- "/Users/madisonparon/Documents/GitHub/james2021/data/fulldata/raw"
out_path <- "/Users/madisonparon/Documents/GitHub/james2021/data/fulldata/clean/FullData_Clea

## ------------------------------------------------------------
## 1. Day 1 and Day 2 file lists, combine files from all nodes and days(many due to branchin
## ------------------------------------------------------------

day1_files <- c(
  "data_exp_250842-v10_task-1ie9.csv",
  "data_exp_250842-v10_task-hxyd.csv",
  "data_exp_250842-v10_task-qglv.csv"
) %>%
  file.path(base_dir, .)

day2_files <- c(
  "data_exp_250842-v10_task-2sj4.csv",
  "data_exp_250842-v10_task-fbey.csv",
```

```
  "data_exp_250842-v10_task-7k22.csv"
) %>%
  file.path(base_dir, .)

## ------------------------------------------------------------
## 2. Item → CBC / neighb mapping
## ------------------------------------------------------------

item_info <- tribble(
  ~item,    ~CBC, ~neighb,
  "peflin", "C",  "none",
  "regby",  "C",  "one",
  "dester", "C",  "many",
  "nusty",  "C",  "many",
  "mowel",  "C",  "many",
  "parung", "C",  "none",
  "pungus", "C",  "one",
  "rafar",  "C",  "one",
  "solly",  "C",  "many",
  "tabric", "C",  "one",
  "tesdar", "C",  "none",
  "vorgal", "C",  "none",
  "wabon",  "C",  "one",
  "ballow", "C",  "many",
  "femod",  "C",  "none"
)

## ------------------------------------------------------------
## 3. Function: convert one file → form-recognition format
## ------------------------------------------------------------

convert_form_recog <- function(path, session_num) {

  raw <- read_csv(path, show_col_types = FALSE)

  resp_rows <- raw %>%
    filter(
      display == "recog-trial",
      `Screen Name` == "response"
    ) %>%
    mutate(
      item = ANSWER,
```

```
      selected = case_when(
        is.na(Response) ~ NA_character_,
        Response == correctResponse ~ ANSWER,
        TRUE ~ foil
      ),

      acc = as.integer(!is.na(selected) & selected == item),

      RT = `Reaction Time`,
      ID = `Participant Public ID`,
      session = session_num
    ) %>%
    left_join(item_info, by = "item")

  resp_rows %>%
    transmute(
      ID,
      CBC,
      item,
      acc,
      RT,
      session,
      neighb
    )
}

## ------------------------------------------------------------
## 4. Apply function to all files and combine
## ------------------------------------------------------------

day1_form <- map_dfr(day1_files, convert_form_recog, session_num = 1)
day2_form <- map_dfr(day2_files, convert_form_recog, session_num = 2)

combined_form <- bind_rows(day1_form, day2_form)

## ------------------------------------------------------------
## 5. Sanity checks
## ------------------------------------------------------------

combined_form %>% count(session)
```

```
# A tibble: 2 x 2
  session     n
    <dbl> <int>
1       1   255
2       2   258
```

```
combined_form %>% count(ID, session)
```

```
# A tibble: 34 x 3
   ID                       session     n
   <chr>                      <dbl> <int>
 1 58a1fe40ea3d11000170d9b9       1    15
 2 58a1fe40ea3d11000170d9b9       2    15
 3 5c04872c55614800012b7cf0       1    15
 4 5c04872c55614800012b7cf0       2    15
 5 5c7dcd11f066ff001568e56a       1    15
 6 5c7dcd11f066ff001568e56a       2    15
 7 5cfc7489e289a80016b6e29f       1    15
 8 5cfc7489e289a80016b6e29f       2    15
 9 60fda57e0916dfe2157c9b00       1    15
10 60fda57e0916dfe2157c9b00       2    15
# i 24 more rows
```

```
combined_form %>% count(item, neighb)
```

```
# A tibble: 15 x 3
   item   neighb     n
   <chr>  <chr>  <int>
 1 ballow many      34
 2 dester many      35
 3 femod  none      34
 4 mowel  many      34
 5 nusty  many      34
 6 parung none      34
 7 peflin none      34
 8 pungus one       34
 9 rafar  one       35
10 regby  one       34
11 solly  many      35
12 tabric one       34
13 tesdar none      34
```

```
14 vorgal none      34
15 wabon  one       34
```

```
## ------------------------------------------------------------
## 6. Save clean combined dataset
## ------------------------------------------------------------

write_csv(combined_form, out_path)
```

3.2.2 Initial analysis

```
library(tidyverse)
library(lme4)
```

```
Loading required package: Matrix
```

```
Attaching package: 'Matrix'
```

```
The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
library(ggpirate)
library(ggplot2)
library(ggpirate)
library(dplyr)
```

```
FR <- read.csv("/Users/madisonparon/Documents/GitHub/james2021/data/fulldata/clean/FullData_C
```

Descriptive statistics

```
FR$neighb <- fct_relevel(FR$neighb, c("none", "one", "many")) # order for intuitive printing

FR %>%
  group_by(session) %>%
  summarise(mean = mean(acc), sd = sd(acc))
```

```
# A tibble: 2 x 3
  session  mean    sd
    <int> <dbl> <dbl>
1       1 0.706 0.457
2       2 0.767 0.423
```

```
FR %>%
  group_by(neighb) %>%
  summarise(mean = mean(acc), sd = sd(acc))
```

```
# A tibble: 3 x 3
  neighb  mean    sd
  <fct>  <dbl> <dbl>
1 none   0.8   0.401
2 one    0.655 0.477
3 many   0.756 0.431
```

```
FR %>%
  group_by(session, neighb) %>%
  summarise(mean = mean(acc), sd = sd(acc))
```

```
`summarise()` has grouped output by 'session'. You can override using the
`.groups` argument.
```

```
# A tibble: 6 x 4
# Groups:   session [2]
  session neighb  mean    sd
    <int> <fct>  <dbl> <dbl>
1       1 none   0.765 0.427
2       1 one    0.624 0.487
3       1 many   0.729 0.447
4       2 none   0.835 0.373
5       2 one    0.686 0.467
6       2 many   0.782 0.416
```

Preparation for analysis

Word Neighbors

```
neighb.contrasts <- cbind(c(-2, 1, 1), c(0,-1,1))
contrasts(FR$neighb) <- neighb.contrasts
contrasts(FR$neighb)
```

```
     [,1] [,2]
none   -2    0
one     1   -1
many    1    1
```

Test Session

```
FR$session <- as.factor(FR$session)
FR$session <- factor(FR$session, levels = c(1, 2), labels = c("Same day", "Next day"))
contrasts(FR$session) <- matrix(c(-1, 1), ncol = 1)
contrasts(FR$session)
```

```
         [,1]
Same day   -1
Next day    1
```

Is performance at above chance levels?

```
# Compute participant means for the first session
pptMeans <- FR %>%
  select(ID, session, acc) %>%
  filter(session == "Same day") %>%
  group_by(ID) %>%
  summarise(mean = mean(acc, na.rm = TRUE))
```

```
# Perform one-sample t-test against chance performance (.5)
t.test(pptMeans$mean, mu=.5, alternative = "greater")
```

```
    One Sample t-test

data:  pptMeans$mean
t = 5.0402, df = 16, p-value = 6.035e-05
alternative hypothesis: true mean is greater than 0.5
95 percent confidence interval:
```

```
 0.6345659       Inf
sample estimates:
mean of x
0.7058824
```

Model building Pruning of fixed effects

```
# Full model with interaction of session and neighb only, drop the random effect with ~0 vari
fix.full <- glmer(
  acc ~ session * neighb + (1 | ID),
  data = FR,
  family = binomial,
  control = glmerControl(optimizer = "bobyqa")
)
```

```
fix.pars <- glmer(acc ~ session*neighb  + (1|ID) + (1|item),
                  data = FR, family = binomial, control = glmerControl(optimizer = "bobyqa"))
anova(fix.pars, fix.full)
```

```
Data: FR
Models:
fix.full: acc ~ session * neighb + (1 | ID)
fix.pars: acc ~ session * neighb + (1 | ID) + (1 | item)
         npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
fix.full    7 568.42 598.10 -277.21    554.42
fix.pars    8 562.77 596.69 -273.38    546.77 7.6492  1    0.00568 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Build random effects structure

```
# Model without random effects for participants
mod.items <- glmer(acc ~ session*neighb + (1|item),
                   data = FR, family = binomial, control = glmerControl(optimizer = "bobyqa"))
anova(mod.items, fix.pars)
```

```
Data: FR
Models:
mod.items: acc ~ session * neighb + (1 | item)
fix.pars: acc ~ session * neighb + (1 | ID) + (1 | item)
```

```
          npar    AIC    BIC  logLik -2*log(L) Chisq Df Pr(>Chisq)
mod.items    7 587.64 617.32 -286.82    573.64
fix.pars     8 562.77 596.69 -273.38    546.77 26.87  1  2.176e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With my limited data, participant variance is estimated ~0. These results indicate that there isn't enough information to reliably estimate participant-level variability. I believe my comparison is structurally identical, but statistically underpowered.

```
# Model without random effects for items
mod.ppts <- glmer(acc ~ session*neighb + (1|ID),
                  data = FR, family = binomial, control = glmerControl(optimizer = "bobyqa"))
anova(mod.ppts, fix.pars)
```

```
Data: FR
Models:
mod.ppts: acc ~ session * neighb + (1 | ID)
fix.pars: acc ~ session * neighb + (1 | ID) + (1 | item)
         npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
mod.ppts    7 568.42 598.10 -277.21    554.42
fix.pars    8 562.77 596.69 -273.38    546.77 7.6492  1    0.00568 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By-participant random slopes

```
# Model with by-participant slopes for the effect of session, an uncorrelated random slope
mod1 <- glmer(
  acc ~ session * neighb + (1 | item) + (1 + session || ID),
  data = FR, family = binomial,
  control = glmerControl(optimizer = "bobyqa")
)
```

```
boundary (singular) fit: see help('isSingular')
```

```
anova(fix.pars, mod1)
```

```
Data: FR
Models:
```

```
fix.pars: acc ~ session * neighb + (1 | ID) + (1 | item)
mod1: acc ~ session * neighb + (1 | item) + (1 + session || ID)
         npar    AIC    BIC  logLik -2*log(L) Chisq Df Pr(>Chisq)
fix.pars    8 562.77 596.69 -273.38    546.77
mod1       11 568.77 615.41 -273.38    546.77     0  3          1
```

```
# Model with by-participant slopes for the effect of neighbor condition
mod2 <- glmer(acc ~ session*neighb + (1+neighb|ID) + (1|item),
              data = FR, family = binomial, control = glmerControl(optimizer = "bobyqa"))
anova(fix.pars, mod2)
```

```
# Model with by-participant slopes for the effects of neighbor condition and test session
mod2a <- glmer(acc ~ session*neighb + (1+neighb+session|ID) + (1|item),
               data = FR, family = binomial, control = glmerControl(optimizer = "bobyqa"))
```

```
anova(mod2, mod2a)
```

By-item random slopes

```
# Model with by-item slopes for the effects of test session
mod3 <- glmer(acc ~ session*neighb + (1+neighb|ID) + (1+session|item),
               data = FR, family = binomial, control = glmerControl(optimizer = "bobyqa"))
```

Final model The final model we arrived at is as follows:

```
summary(mod2)
```

Plot Graphs

```
##Compute participant condition means
cond.means <- FR %>%
  group_by(ID, session, neighb) %>%
  summarise(mean = mean(acc, na.rm = TRUE))
```

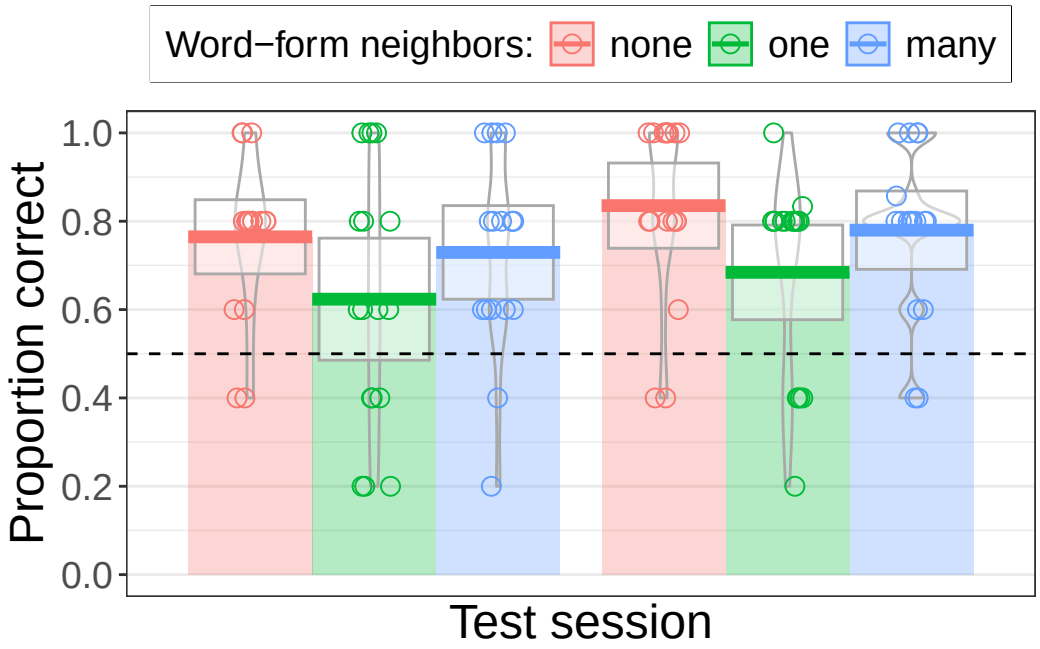
```
`summarise()` has grouped output by 'ID', 'session'. You can override using the
`.groups` argument.
```

```
cond.means$session <- as.factor(cond.means$session)
```

```r
formGraph <- ggplot(data = cond.means, aes(x = session, y = mean)) +
  theme_bw() +
  geom_pirate(aes(colour = neighb, fill = neighb),
              show.legend   = TRUE,
              points_params = list(alpha = 1, size = 3),
              lines_params  = list(size = 0.9)) +
  scale_color_discrete(name = "Word-form neighbors:") +
  scale_fill_discrete(name = "Word-form neighbors:") +
  scale_x_discrete(
    name   = "Test session",
    breaks = c("1", "2"),           # use c("1","2","3") if you have 3 sessions
    labels = c("Same day", "Next day")
    # labels = c("Same day","Next day","Week later") for 3 sessions
  ) +
  scale_y_continuous(
  name   = "Proportion correct",
  breaks = seq(0, 1, by = 0.2),
  limits = c(0, 1)
  ) +
  theme(
    legend.position       = "top",
    legend.title          = element_text(size = 14),
    legend.text           = element_text(size = 14),
    legend.box.background = element_rect(colour = "black"),
    strip.text            = element_text(size = 20),
    axis.title            = element_text(size = 18),
    axis.text             = element_text(size = 14)
  ) +
  geom_hline(yintercept = 0.50, linetype = "dashed")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```r
print(formGraph)
```




```
# Save as a ggplot object
saveRDS(formGraph, file = "formGraph.rds")
```

R Version information

```
sessionInfo()
```

```
R version 4.5.1 (2025-06-13)
Platform: aarch64-apple-darwin20
Running under: macOS Tahoe 26.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib;

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: America/Los_Angeles
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base
```

```
other attached packages:
 [1] ggpirate_0.1.2  lme4_1.1-37     Matrix_1.7-3    lubridate_1.9.4
 [5] forcats_1.0.1   stringr_1.5.2   dplyr_1.1.4     purrr_1.1.0
 [9] readr_2.1.5     tidyr_1.3.1     tibble_3.3.0    ggplot2_4.0.0
[13] tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] utf8_1.2.6          generics_0.1.4      stringi_1.8.7       lattice_0.22-7
 [5] hms_1.1.3           digest_0.6.37       magrittr_2.0.4      evaluate_1.0.5
 [9] grid_4.5.1          timechange_0.3.0    RColorBrewer_1.1-3  fastmap_1.2.0
[13] jsonlite_2.0.0      tinytex_0.57        scales_1.4.0        Rdpack_2.6.4
[17] reformulas_0.4.2    cli_3.6.5           rlang_1.1.6         crayon_1.5.3
[21] rbibutils_2.4       bit64_4.6.0-1       splines_4.5.1       withr_3.0.2
[25] yaml_2.3.10         tools_4.5.1         parallel_4.5.1      tzdb_0.5.0
[29] nloptr_2.2.1        minqa_1.2.8         boot_1.3-31         vctrs_0.6.5
[33] R6_2.6.1            lifecycle_1.0.4     bit_4.6.0           MASS_7.3-65
[37] vroom_1.6.6         pkgconfig_2.0.3     pillar_1.11.1       gtable_0.3.6
[41] Rcpp_1.1.0          glue_1.8.0          xfun_0.54           tidyselect_1.2.1
[45] rstudioapi_0.17.1   knitr_1.50          farver_2.1.2        nlme_3.1-168
[49] htmltools_0.5.8.1   rmarkdown_2.30      compiler_4.5.1      S7_0.2.0
```

3.3 Exploratory analyses

Any follow-up analyses desired (not required).

4 Results/Discussion

4.1 Descriptive Results

The figure I was able to create from my data shows mean form-recognition accuracy as a function of test session (Same day vs. Next day) and phonological neighborhood condition (none, one, many). Performance was above chance in both sessions and across all neighborhood conditions. Mean accuracy increased from the same-day test (M = 0.71, SD = 0.46) to the next-day test (M = 0.77, SD = 0.42), consistent with an overall benefit of offline consolidation. Across sessions, words with no phonological neighbors were recognized most accurately (M = 0.80), followed by many-neighbor words (M = 0.76), with one-neighbor words showing the lowest accuracy (M = 0.66).

A one-sample t-test on participant-level means from the same-day test confirmed that performance was significantly above chance (M = 0.71), t(16) = 5.04, p < .001, indicating reliable learning from incidental exposure.

4.2 Summary of Replication Attempt

This study sought to replicate the adult form-recognition results of Experiment 2 from James et al. (2021), examining how phonological neighborhood structure influences incidental word learning and its consolidation over time. The replication was partially successful. Participants reliably learned novel word forms from incidental story exposure, and performance was significantly above chance. Descriptively, accuracy improved from the same-day test to the next-day test, consistent with consolidation-related gains.

However, the replication did not fully reproduce the original inferential pattern. In particular, the present data did not support complex random-effects structures or yield strong evidence for interactions between session and neighborhood condition.

4.3 Mixed-Effects Modeling

Form-recognition accuracy was analyzed using mixed-effects logistic regression models with fixed effects of test session, phonological neighborhood condition, and their interaction. Orthogonal contrasts were used to compare (i) words with no neighbors versus words with neighbors, and (ii) words with one versus many neighbors.

4.3.1 Random Intercepts

Comparisons between models with and without random intercepts indicated that including random effects for both participants and items significantly improved model fit relative to models omitting either source of variance (both ps < .01). This supports the inclusion of crossed random intercepts for participants and items in subsequent models.

4.3.2 Random Slopes

Following the forward best-path approach used in the original study, random slopes were added incrementally. Adding a by-participant random slope for test session did not improve model fit relative to the intercepts-only model, $^{2}(3) = 0.01$, p = .999, and resulted in singular fits. Similarly, more complex random-slope structures were not supported by the data.

As a result, the final model included random intercepts for participants and items, but no random slopes. This contrasts with the original study, which supported by-participant random slopes for session effects.

4.4 Comparison to Original Findings/Commentary

The overall pattern of results partially replicates the findings of James et al. (2021). As in the original study, participants demonstrated above-chance learning from incidental exposure, and descriptive results suggest improved performance after a delay. However, unlike the original study, the present replication did not support a more complex random-effects structure, nor did it provide sufficient evidence to detect interactions between session and neighborhood condition.

Importantly, the present study included only two test sessions (same day and next day), whereas the original study included a third delayed session one week later. The absence of this longer delay likely reduced sensitivity to consolidation-related changes that were central to the original theoretical claims.

4.5 Interpreting Differences from the Original Study

Several methodological differences likely contributed to these discrepancies.

First, the present replication included only two test sessions rather than three. The original study demonstrated that neighborhood effects became more pronounced after longer consolidation intervals, particularly at the one-week delay. By omitting this session, the present study may have attenuated precisely the effects most diagnostic of consolidation-based lexical integration.

Second, the sample size was substantially smaller than in the original study, limiting statistical power. This is reflected in frequent singular model fits and the inability to justify random slopes, despite following an identical modeling strategy. These issues suggest that the data contained insufficient information to reliably estimate participant-level variability in learning trajectories.

Third, although the procedural framework closely followed the original study—including use of the same experimental platform (Gorilla), stimuli, and task structure—the use of a U.S.-based Prolific sample rather than a U.K.-based sample may have introduced additional variability in phonological familiarity or response strategies.

4.6 Implications and Conclusions

Despite these limitations, the replication provides converging evidence that adults can incidentally acquire novel word forms from narrative exposure and that performance improves after a short delay. The failure to fully replicate the original random-effects structure and interaction patterns appears most consistent with reduced power and the absence of a longer consolidation interval, rather than a substantive failure of the theoretical account.

Taken together, these findings suggest that phonological neighborhood effects on lexical consolidation are likely robust, but their detectability depends critically on sufficient longitudinal sampling and statistical power. Future replication efforts should prioritize longer retention intervals and larger samples to more fully evaluate consolidation-based predictions.