Morphological predictability - HS: 1.2

EDA - proficiency/use

Joseph V. Casillas

2023-04-04

Table of contents

Setup

Load libraries and helper functions:

```
source(here::here("scripts", "00_libraries.R"))
source(here::here("scripts", "01_helpers.R"))
```

Load proficiency and use data:

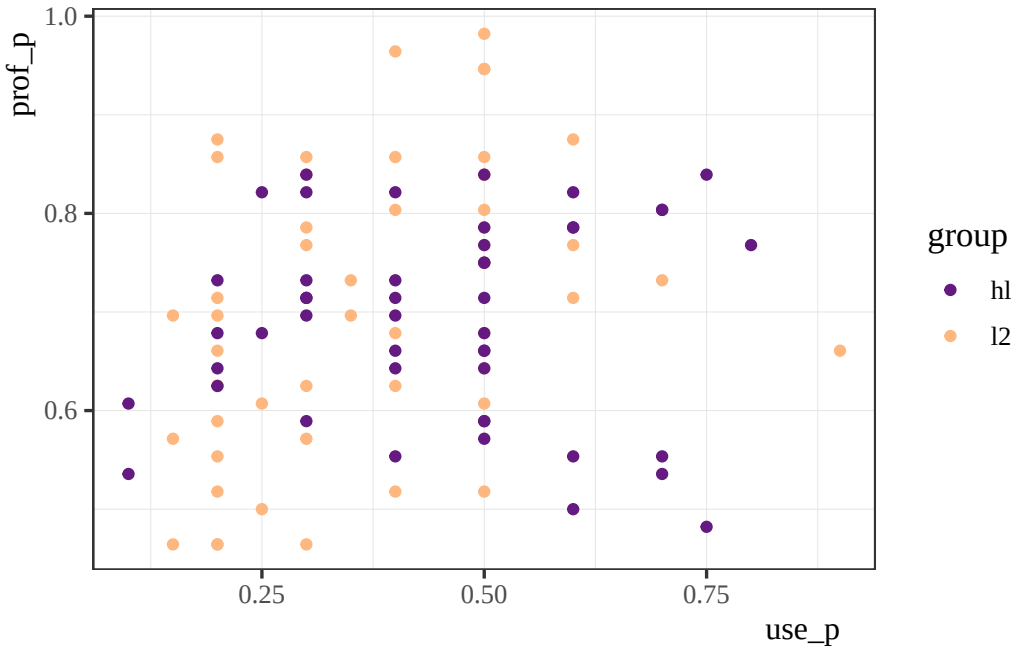
```
prof_use_temp <- read_xlsx(
  here("data", "raw",  "HS_L2_used_blc_V2.xlsx")
  ) %>%
  rename(id = Participant, prof = dele, use = `span_week_%`)
```

Next, we tidy the data and convert to z-scores.

```
# Create heritage and L2 groups
# Standardize proficiency
prof_use <- prof_use_temp %>%
  mutate(group = case_when(
    Type == "advanced_heritage" ~ "hl",
    Type == "intermediate_heritage" ~ "hl",
    TRUE ~ "l2"),
    prof_p = prof / 56,
    use_p = use / 100,
    prof_perc = prof_p * 100,
    use_perc = use_p * 100,
    prof_z = (prof_perc - mean(prof_perc)) / sd(prof_perc),
    use_z = (use_perc - mean(use_perc)) / sd(use_perc)
  )
```

Basic plot to check distributions.

```r
prof_use %>%
  ggplot(., aes(x = use_p, y = prof_p, color = group)) +
    geom_point() +
    scale_color_viridis_d(option = "A", begin = 0.3, end = 0.85)
```



The data look pretty evenly distributed, though perhaps there is more variability in the L2 group. Now we will look at some descriptives in order to get a sense of whether or not there are any individuals that might make the groups uneven.

```r
# Check mean (SD) for each TYPE
prof_use %>%
  group_by(Type) %>%
  summarize(across(c("prof_p", "use_p"),
    list(mean = mean, sd = sd)))

# Might be able to balance groups by removing intermediate
# heritage with high use
prof_use %>%
  filter(Type == "intermediate_heritage") %>%
  arrange(desc(use_p)) %>%
  select(id, prof_p, use_p)
```

```r
# There are three in the 70+ % range
# We'll start there
prof_use %>%
  filter(!(id %in% c("ihs24", "ihs13", "ihs18"))) %>%
  group_by(Type) %>%
  summarize(across(c("prof_p", "use_p"),
    list(mean = mean, sd = sd)))

# They are closer now. We can also take out a few of the lower L2 use
prof_use %>%
  filter(Type == "intermediate_l2") %>%
  arrange(use_p) %>%
  select(id, prof_p, use_p)

# There are three in the 15% range
# We'll start there
prof_use %>%
  filter(!(id %in% c("ihs24", "ihs13", "ihs18")),
         !(id %in% c("ies09", "ies26", "ies29"))) %>%
  group_by(Type) %>%
  summarize(across(c("prof_p", "use_p"),
    list(mean = mean, sd = sd)))

prof_use %>%
  filter(Type == "advanced_heritage") %>%
  arrange(desc(use_p)) %>%
  select(id, prof_p, use_p)


prof_use %>%   filter(!(id %in% c("ihs24", "ihs13", "ihs18")),
          !(id %in% c("ahs12", "ahs03")),
          !(id %in% c("ies09", "ies26", "ies29"))) %>%
  group_by(group) %>%
  summarize(across(c("prof_p", "use_p"),
    list(mean = mean, sd = sd)))

# IDs to drop
remove <- c(
  "ihs24", "ihs13", "ihs18",
  "ahs12", "ahs03",
  "ies09", "ies26", "ies29"
  )

prof_use_final <- prof_use %>%
  filter(!(id %in% remove)) %>%
```

```
    mutate(prof_z = (prof_perc - mean(prof_perc)) / sd(prof_perc),
      use_z = (use_perc - mean(use_perc)) / sd(use_perc))
```

```
# A tibble: 4 x 5
  Type                  prof_p_mean prof_p_sd use_p_mean use_p_sd
  <chr>                       <dbl>     <dbl>      <dbl>    <dbl>
1 advanced_heritage           0.776    0.0450      0.474    0.169
2 advanced_l2                 0.823    0.0775      0.437    0.134
3 intermediate_heritage       0.609    0.0633      0.421    0.185
4 intermediate_l2             0.591    0.0780      0.307    0.167
# A tibble: 24 x 3
   id    prof_p use_p
   <chr>  <dbl> <dbl>
 1 ihs24  0.482  0.75
 2 ihs13  0.536  0.7
 3 ihs18  0.554  0.7
 4 ihs06  0.5    0.6
 5 ihs27  0.554  0.6
 6 ihs02  0.661  0.5
 7 ihs04  0.589  0.5
 8 ihs09  0.589  0.5
 9 ihs10  0.661  0.5
10 ihs11  0.643  0.5
# i 14 more rows
# A tibble: 4 x 5
  Type                  prof_p_mean prof_p_sd use_p_mean use_p_sd
  <chr>                       <dbl>     <dbl>      <dbl>    <dbl>
1 advanced_heritage           0.776    0.0450      0.474    0.169
2 advanced_l2                 0.823    0.0775      0.437    0.134
3 intermediate_heritage       0.622    0.0566      0.379    0.155
4 intermediate_l2             0.591    0.0780      0.307    0.167
# A tibble: 27 x 3
   id    prof_p use_p
   <chr>  <dbl> <dbl>
 1 ies09  0.571  0.15
 2 ies26  0.696  0.15
 3 ies29  0.464  0.15
 4 ies05  0.625  0.2
 5 ies06  0.661  0.2
 6 ies07  0.625  0.2
 7 ies12  0.464  0.2
 8 ies13  0.696  0.2
 9 ies20  0.625  0.2
10 ies21  0.518  0.2
# i 17 more rows
```

```
# A tibble: 4 x 5
  Type                  prof_p_mean prof_p_sd use_p_mean use_p_sd
  <chr>                       <dbl>     <dbl>      <dbl>    <dbl>
1 advanced_heritage           0.776    0.0450      0.474    0.169
2 advanced_l2                 0.823    0.0775      0.437    0.134
3 intermediate_heritage       0.622    0.0566      0.379    0.155
4 intermediate_l2             0.592    0.0753      0.327    0.167
# A tibble: 23 x 3
   id    prof_p use_p
   <chr>  <dbl> <dbl>
 1 ahs12  0.768  0.8
 2 ahs03  0.839  0.75
 3 ahs15  0.804  0.7
 4 ahs23  0.804  0.7
 5 ahs04  0.821  0.6
 6 ahs11  0.786  0.6
 7 ahs16  0.786  0.6
 8 ahs02  0.768  0.5
 9 ahs06  0.839  0.5
10 ahs10  0.786  0.5
# i 13 more rows
# A tibble: 2 x 5
  group prof_p_mean prof_p_sd use_p_mean use_p_sd
  <chr>       <dbl>     <dbl>      <dbl>    <dbl>
1 hl          0.698    0.0921      0.412    0.153
2 l2          0.712    0.139       0.384    0.159
```

We now have a few candidates. We will do a few formal tests to look at equivalence in proficiency and use with and without them.

```
#
# Fit models
#

# model formula
model_formula <- bf(
  prop ~ group,
  phi  ~ group
  )

# Fit proficiency model with all data
prof_with <- brm(
  formula = model_formula,
  family = Beta(),
  data = mutate(prof_use, prop = prof / 56),
  file = here("models", "eda_prof_use", "prof_with")
```

```
  )

# Fit use model with all data
use_with <- brm(
  formula = model_formula,
  family = Beta(),
  data = mutate(prof_use, prop = use / 100),
  file = here("models", "eda_prof_use", "use_with")
  )

# Fit proficiency model remove a few participants
prof_wo <- brm(
  formula = model_formula,
  family = Beta(),
  data = mutate(prof_use_final, prop = prof / 56),
  file = here("models", "eda_prof_use", "prof_wo")
  )

# Fit use model remove a few participants
use_wo <- brm(
  formula = model_formula,
  family = Beta(),
  data = mutate(prof_use_final, prop = use / 100),
  file = here("models", "eda_prof_use", "use_wo")
  )
```

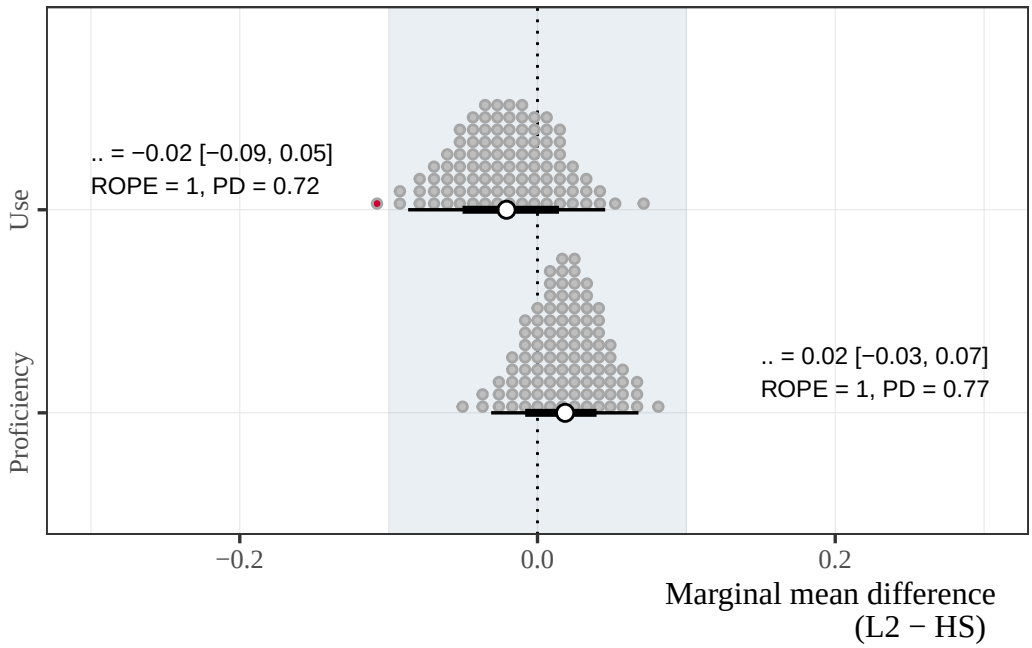
Now we can check the marginal mean estimates to see how the groups compare.

```
# Get marginal mean estimates for prof with ids removed
prof_contrast <- prof_wo %>%
  emmeans(spec = ~ group,
    at = list(group = c("l2", "hl")),
    epred = TRUE) %>%
  contrast(method = "pairwise") %>%
  gather_emmeans_draws() %>%
  mutate(metric = "prof")

# Get marginal mean estimates for use with ids removed
use_contrast <- use_wo %>%
  emmeans(spec = ~ group,
    at = list(group = c("l2", "hl")),
    epred = TRUE) %>%
  contrast(method = "pairwise") %>%
  gather_emmeans_draws() %>%
  mutate(metric = "use")
```

```r
use_prof_desc <- bind_rows(
  bind_cols(x = 0.15, y = 1.2,
    metric = "Proficiency",
    median_hdi(prof_contrast, .value),
    ROPE = rope(prof_contrast$.value, ci_method = "HDI")$ROPE_Percentage,
    PD = p_direction(prof_contrast$.value)[[1]]
    ),
  bind_cols(x = -0.3, y = 2.2,
    metric = "Use",
    median_hdi(use_contrast, .value),
    ROPE = rope(use_contrast$.value, ci_method = "HDI")$ROPE_Percentage,
    PD = p_direction(use_contrast$.value)[[1]]
    )
  ) %>%
  mutate(across(c(".value", ".lower", ".upper", "PD"),
    ~ round(.x, digits = 2))) %>%
  transmute(x, y, metric, contrast,
    out = glue("  = {.value} [{.lower}, {.upper}]\nROPE = {ROPE}, PD = {PD}"))

bind_rows(prof_contrast, use_contrast) %>%
  mutate(ylab = if_else(metric == "use", "Use", "Proficiency")) %>%
  ggplot(., aes(x = .value, y = ylab, fill = stat(abs(x) > 0.1))) +
    geom_tile(data = tibble(
      ylab = "Proficiency", value = 0, lower = -0.1, upper = 0.1
      ),
      aes(x = value, width = lower - upper),
      alpha = 0.1, height = Inf, fill = "#31688EFF") +
    geom_vline(xintercept = 0, lty = 3, size = 0.5) +
    stat_dotsinterval(quantiles = 100, point_interval = "median_hdi",
      show.legend = F, pch = 21, point_fill = "white") +
    scale_fill_manual(values = c("grey", "#cc0033")) +
    coord_cartesian(xlim = c(-0.3, 0.3)) +
    geom_text(data = use_prof_desc, aes(x = x, y = y, label = out),
      hjust = 0, size = 3) +
    labs(y = NULL, x = "Marginal mean difference\n(L2 - HS)") +
    theme(axis.text.y = element_text(angle = 90, hjust = 0.5))
```




```
n_groups <- prof_use_final %>%
  mutate(group = str_replace_all(group, "hl", "HS")) %>%
  group_by(group) %>%
  summarize(n = n_distinct(id))

tab_prof_use_desc <- prof_use_final %>%
  mutate(group = str_replace_all(group, "hl", "HS")) %>%
  group_by(group) %>%
  summarize(across(c("prof_p", "use_p"),
    list(mean = mean, sd = sd))) %>%
  mutate_if(is.numeric, specify_decimal, k = 2) %>%
  left_join(., n_groups, by = "group") %>%
  mutate(
    Proficiency = glue("{prof_p_mean} ({prof_p_sd})"),
    Proficiency_prose = glue("{prof_p_mean} (SD = {prof_p_sd})"),
    Use = glue("{use_p_mean} ({use_p_sd})"),
    Use_prose = glue("{use_p_mean} (SD = {use_p_sd})"),
    ) %>%
  transmute(Group = group, n, Proficiency, Use, Proficiency_prose, Use_prose) %>%
  mutate(Group = str_to_upper(Group)) %>%
  write_csv(here("tables", "tab_prof_use_desc.csv"))
```

```r
tab_prof_use_desc %>%
  select(-c("Proficiency_prose", "Use_prose")) %>%
  knitr::kable(format = "pandoc",
    align = c("l", "r", "r", "r"),
    caption = "Average use and proficiency for L2 and heritage learners.")
```

Table 1: Average use and proficiency for L2 and heritage learners.

Group	n	Proficiency	Use
HS	42	0.70 (0.09)	0.41 (0.15)
L2	50	0.71 (0.14)	0.38 (0.16)

```r
tab_use_prof_desc_mod <- use_prof_desc %>%
  mutate(contrast = str_replace_all(contrast, "hl", "hs")) %>%
  select(metric:out) %>%
  separate(out, into = c("Estimate", "ROPE"), sep = "\n") %>%
  separate(ROPE, into = c("ROPE", "PD"), sep = ",") %>%
  mutate(contrast = str_to_upper(contrast),
    Estimate = str_remove(Estimate, "  = "),
    Estimate = str_replace_all(Estimate, "-", "\U2212"),
    PD = str_remove(PD, "PD = "),
    ROPE = str_remove(ROPE, "ROPE = "),
    PD = str_remove(PD, " ")) %>%
  select(Metric = metric, Contrast = contrast, Estimate, ROPE, PD) %>%
  write_csv(here("tables", "tab_prof_use_mod.csv"))

tab_use_prof_desc_mod %>%
  knitr::kable(format = "pandoc",
    align = c("l", "l", "r", "r", "r"),
    caption = "Marginal mean difference beween L2 and heritage leareners for use and proficiency."
```

Table 2: Marginal mean difference beween L2 and heritage leareners for use and proficiency.

Metric	Contrast	Estimate	ROPE	PD
Proficiency	L2 - HS	0.02 [−0.03, 0.07]	1	0.77
Use	L2 - HS	−0.02 [−0.09, 0.05]	1	0.72

```r
prof_use %>%
  mutate(remove = if_else(id %in% remove, 1, 0)) %>%
  write_csv(here("data", "tidy", "prof_use.csv"))
```