The data comes from a published article by [Vesker et al. (2018)](https://www.tandfonline.com/doi/full/10.1080/17405629.2017.1287073). Briefly, this study involved having adults and 9-year-old children rate a series of 48 faces on two dimensions of emotion, valence (positive vs. negative) and arousal (activated vs. deactivated) (see my previous [post](https://willhipson.netlify.com/post/circumplex/circumplex/) for more info on valence and arousal). The authors made some interesting observations about differences in childrens’ and adult’s ratings of these facial expressions.

However, absent from the writeup was a discussion about how reliable these ratings are. We might wonder about the extent to which people agree on the valence or arousal of a face and whether this varies between children and adults. Here, I tackle the issue of intraclass correlation (ICC) using the dataset published by Vesker et al. (2018). The data itself is openly accessible [here](https://zenodo.org/record/293008#.XFXlr1xKhPZ).

First, I’ll load up the *tidyverse* and *readxl* packages, which will help with the data cleaning.

library(readxl)

library(tidyverse)

options(digits = 3, scipen = -2)

Upon downloading the data, we’re immediately presented with an issue: it’s an xlsx document (Excel) containing multiple sheets, with each sheet representing a “condition” (Child vs. Adult) and (Valence vs. Arousal). On Stack Overflow, I found a useful function for reading in multiple sheets ([see here for original post](https://stackoverflow.com/questions/12945687/read-all-worksheets-in-an-excel-workbook-into-an-r-list-with-data-frames)).

read\_excel\_allsheets <- function(filename, tibble = TRUE) {

sheets <- readxl::excel\_sheets(filename)

x <- lapply(sheets, function(x) readxl::read\_excel(filename, sheet = x))

if(!tibble) x <- lapply(x, as.data.frame)

names(x) <- sheets

x

}

Now, I’ll read in the data and extract each sheet. I’m using VA to refer to valence and AR for arousal.

faces <- read\_excel\_allsheets("C:/Users/wille/Downloads/adults and 9yo all AR and VAL face ratings zenodo.xlsx")

VA\_adult\_raw <- faces$`VAL adult faces`

VA\_child\_raw <- faces$`Val 9yo faces`

AR\_adult\_raw <- faces$`AR adult faces`

AR\_child\_raw <- faces$`AR 9yo faces`

Let’s get a look at one of these datasets.

head(VA\_adult\_raw[, 1:20]) #Limiting preview to n = 20

## # A tibble: 6 x 20

## ..1 ..2 ..3 ..4 ..5 ..6 ..7 ..8 ..9 ..10 ..11 ..12

##

## 1 dl sa ~ 4 4 3 4 2 3 4 3 3 5 3

## 2 dl sa ~ 4 4 3 3 2 3 3 3 3 4 4

## 3 lp sa ~ 2 3 4 3 3 3 2 3 3 3 2

## 4 lp sa ~ 4 3 2 3 3 3 4 2 2 4 2

## 5 ma sa ~ 1 1 1 2 1 2 1 1 2 3 2

## 6 md sa ~ 2 1 1 2 2 3 1 1 2 3 2

## # ... with 8 more variables: ..13 , ..14 , ..15 ,

## # ..16 , ..17 , ..18 , ..19 , ..20

It’s not immediately apparent what is being displayed here because the columns aren’t labeled. The article tells us that participants rated 48 faces, so based on the dimensions we can assume that each row is a face and each column is a participant who rated that face. Admittedly, it’s a less intuitive way of representing the data, but its actually ideal for computing ICCs.

Still, there’s a lot of data cleaning and wrangling to be done here. First, we have some rows and columns that aren’t relevant, so we’ll get rid of those. Of note, *dplyr*’s lesser known *slice* function is helpful for identifying which rows we want to keep.

VA\_adult <- VA\_adult\_raw %>%

slice(1:48) %>%

select(-1, -mean, -SD, -`M mean`, -`F mean`, -`dist from 4`, -`0`, -aa, -Valence) %>%

mutate(face = row\_number()) %>%

select(face, everything())

Now, to make the columns more intuitive, we’ll label them properly.

colnames(VA\_adult)[2:161] <- paste("rater", 1:160)

**Plotting the Scores**

We may want to plot the data to get a sense of the variability around raters’ labeling of valence across the 48 faces. Our data is currently in wide form, but we need to set it up such that all of the ratings are in one column. This is where *reshape2* comes into action.

library(reshape2)

The function *melt* will take our wide dataset and make it long. We supply it with an *id.vars* which tells it which of our original columns we want to stay as a column. Then it takes all of the other columns and condenses them into one variable.

VA\_adult\_melt <- VA\_adult %>%

melt(id.vars = "face", value.name = "valence", variable.name = "rater")

head(VA\_adult\_melt)

## face rater valence

## 1 1 rater 1 4

## 2 2 rater 1 4

## 3 3 rater 1 2

## 4 4 rater 1 4

## 5 5 rater 1 1

## 6 6 rater 1 2

Now we’ll turn this into a line graph with each line representing an individual rater’s valence ratings for each of the 48 faces. It will be crowded, but that’s OK. We just want to see if the lines cluster around each other or not.

VA\_adult\_melt %>%

ggplot(aes(face, valence, group = rater, color = rater)) +

geom\_line(size = .8, alpha = .5) +

scale\_x\_discrete(limits = 1:48) +

geom\_vline(xintercept = 24.5, size = 1.5, color = "red") +

labs(x = "Face", y = "Valence (higher = more positive)",

title = "Adult Valence Ratings for 48 Faces",

subtitle = "Red line indicates where faces become positive") +

theme\_bw() +

theme(legend.position = "none")

There’s a clear division between the positive and negative faces. It seems that there’s strong agreement among adults as to what constitutes a positive or negative face.

What if we looked at the distributions of the ratings for each face?

VA\_adult\_melt %>%

group\_by(face) %>%

summarize(mean = mean(valence),

sd = sd(valence)) %>%

ungroup() %>%

ggplot(aes(face, mean)) +

geom\_errorbar(aes(ymin = mean - 1.96\*(sd/sqrt(ncol(VA\_adult))),

ymax = mean + 1.96\*(sd/sqrt(ncol(VA\_adult))))) +

geom\_point(size = 2) +

scale\_x\_discrete(limits = 1:48) +

geom\_vline(xintercept = 24.5, color = "red") +

labs(x = "Face", y = "Valence (higher = more positive)",

title = "Adult - Average Valence Ratings for 48 Faces",

subtitle = "Red line indicates where faces become positive; error bars = 95% CI") +

theme\_bw()

It tells the same story, but it’s a more polished figure. Notice how the error bars for the 95% confidence interval around the mean are quite small.

**Cleaning the Remaining Datasets**

First, we’ll look at the dataset for 9-year-olds’ ratings of valence. Note that there are a few modifications to the script due to idiosyncracies with the original datasets.

VA\_child <- VA\_child\_raw %>%

slice(2:49) %>%

select(-1, -`average child ratings`, -33, -code,

-`pic name`, -emotion, -`Child M`, -`Child F`, -sex, -Valence) %>%

mutate(face = row\_number()) %>%

select(face, everything())

colnames(VA\_child)[2:31] <- paste("rater", 1:30)

VA\_child <- tbl\_df(lapply(VA\_child, function(x){ #Need to use a function to convert to numeric

as.numeric(as.character(x))

}))

VA\_child\_melt <- VA\_child %>%

melt(id.vars = "face", value.name = "valence", variable.name = "rater")

VA\_child\_melt %>%

ggplot(aes(face, valence, group = rater, color = rater)) +

geom\_line(size = .8, alpha = .5) +

scale\_x\_discrete(limits = 1:48) +

geom\_vline(xintercept = 24.5, size = 1.5, color = "red") +

labs(x = "Face", y = "Valence (higher = more positive)",

title = "Child Valence Ratings for 48 Faces",

subtitle = "Red line indicates where faces become positive") +

theme\_bw() +

theme(legend.position = "none")

VA\_child\_melt %>%

group\_by(face) %>%

summarize(mean = mean(valence, na.rm = TRUE),

sd = sd(valence, na.rm = TRUE)) %>%

ungroup() %>%

ggplot(aes(face, mean)) +

geom\_errorbar(aes(ymin = mean - 1.96\*(sd/sqrt(ncol(VA\_child))),

ymax = mean + 1.96\*(sd/sqrt(ncol(VA\_child))))) +

geom\_point(size = 2) +

scale\_x\_discrete(limits = 1:48) +

geom\_vline(xintercept = 24.5, color = "red") +

labs(x = "Face", y = "Valence (higher = more positive)",

title = "Child - Average Valence Ratings for 48 Faces",

subtitle = "Red line indicates where faces become positive; error bars = 95% CI") +

theme\_bw()

The results are similar to those from the adults. We can’t trust the wider confidence intervals to tell us about reliability though, because there are far fewer child raters than adult raters.

**Repeating the Procedure for Ratings of Arousal**

Finally, we repeat the analysis for measures of arousal, starting with adults then with children.

AR\_adult <- AR\_adult\_raw %>%

slice(1:48) %>%

select(-1, -43, -SD, -`M mean`, -`F mean`, -mean, -`0`, -Valence) %>%

mutate(face = row\_number()) %>%

select(face, everything())

colnames(AR\_adult)[2:42] <- paste("rater", 1:41)

AR\_adult\_melt <- AR\_adult %>%

melt(id.vars = "face", value.name = "arousal", variable.name = "rater")

AR\_adult\_melt %>%

ggplot(aes(face, arousal, group = rater, color = rater)) +

geom\_line(size = .8, alpha = .5) +

scale\_x\_discrete(limits = 1:48) +

geom\_vline(xintercept = 24.5, size = 1.5, color = "red") +

labs(x = "Face", y = "Arousal (higher = more activated)",

title = "Adult Arousal Ratings for 48 Faces",

subtitle = "Red line indicates where faces become positive") +

theme\_bw() +

theme(legend.position = "none")

There seems to be much less consensus for ratings of arousal. We do notice that there is no differentiation between positive and negative faces – this is good because the theory suggests that arousal is independent of valence. Someone can be positively aroused (e.g., excited) or negatively aroused (e.g., stressed). However, if there was high consensus we would still see the lines converging. Instead, they’re all over the place.

AR\_adult\_melt %>%

group\_by(face) %>%

summarize(mean = mean(arousal),

sd = sd(arousal)) %>%

ungroup() %>%

ggplot(aes(face, mean)) +

geom\_errorbar(aes(ymin = mean - 1.96\*(sd/sqrt(ncol(AR\_adult))),

ymax = mean + 1.96\*(sd/sqrt(ncol(AR\_adult))))) +

geom\_point(size = 2) +

scale\_x\_discrete(limits = 1:48) +

geom\_vline(xintercept = 24.5, color = "red") +

labs(x = "Face", y = "Arousal (higher = more activated)",

title = "Adult - Average Arousal Ratings for 48 Faces",

subtitle = "Red line indicates where faces become positive; error bars = 95% CI") +

theme\_bw()

Confidence intervals are much wider too, but again, we have a smaller sample size so that adds some uncertainty. Still, it seems like adults have difficulty agreeing on ratings of arousal compared to ratings of valence. Let’s go back to 9-year-olds.

AR\_child <- AR\_child\_raw %>%

slice(2:49) %>%

select(-`photo ID`, -child, -`adult ratings PELL`, -`Photo ID`,

-30, -image, -emotion, -`m child`, -`f child`, -Valence) %>%

mutate(face = row\_number()) %>%

select(face, everything())

colnames(AR\_child)[2:31] <- paste("rater", 1:30)

AR\_child <- tbl\_df(lapply(AR\_child, function(x){ #Need to use a function to convert to numeric

as.numeric(as.character(x)) #Note: There is one missing value from original dataset

}))

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

AR\_child\_melt <- AR\_child %>%

melt(id.vars = "face", value.name = "arousal", variable.name = "rater")

AR\_child\_melt %>%

ggplot(aes(face, arousal, group = rater, color = rater)) +

geom\_line(size = .8, alpha = .5) +

scale\_x\_discrete(limits = 1:48) +

geom\_vline(xintercept = 24.5, size = 1.5, color = "red") +

labs(x = "Face", y = "Arousal (higher = more activated)",

title = "Child Arousal Ratings for 48 Faces",

subtitle = "Red line indicates where faces become positive") +

theme\_bw() +

theme(legend.position = "none")

## Warning: Removed 48 rows containing missing values (geom\_path).

AR\_child\_melt %>%

group\_by(face) %>%

summarize(mean = mean(arousal, na.rm = TRUE),

sd = sd(arousal, na.rm = TRUE)) %>%

ungroup() %>%

ggplot(aes(face, mean)) +

geom\_errorbar(aes(ymin = mean - 1.96\*(sd/sqrt(ncol(AR\_child))),

ymax = mean + 1.96\*(sd/sqrt(ncol(AR\_child))))) +

geom\_point(size = 2) +

scale\_x\_discrete(limits = 1:48) +

geom\_vline(xintercept = 24.5, color = "red") +

labs(x = "Face", y = "Arousal (higher = more activated)",

title = "Child - Average Arousal Ratings for 48 Faces",

subtitle = "Red line indicates where faces become positive; error bars = 95% CI") +

theme\_bw()

Results look similar for children. I won’t spend much time discussing mean differences in valence and arousal between children and adults – the original article expands on this. However, I am interested in the variability in ratings of arousal vs. valence.

**Quantifying Interrater Agreement**

So far, we’ve created a series of plots showing the variability in childrens’ and adult’s ratings of emotional facial expressions. We get a sense that both children and adults reliably label faces as positive or negative, but they struggle to agree on arousal. Although this is apparent from the plots, we may want to test this more formally. This is actually very important because our estimates of variability (e.g., 95% CI) are sensitive to sample size, which varies by adults and children in this dataset.

**Intra-correlation coefficient (ICC)**

The ICC is an index of reliability or agreement for continuous ratings. ICCs range from 0 (no agreement) to 1 (perfect agreement). We will use ICC to quantify agreement on ratings of emotional facial expressions, but ICC is applicable to other situations, such as quantifying heritability or assessing items in a test bank. Here, we will calculate four ICCs: (1) Adult ratings of Valence, (2) Child ratings of Valence, (3) Adults ratings of Arousal, and (4) Child ratings of Arousal.

Shrout and Fleiss (1979), and later McGraw and Wong (1996), describe several different calculations for ICC that depend on the characteristics of the sample. In our case, we will use a two-way random model for single measurements to quantify absolute agreement, also known as ICC2.

Two way random, single measures, absolute (ICC2):

\[\rho = \frac{\sigma^2\_r}{\sigma^2\_r + \sigma^2\_c + \sigma^2\_e}\]

Where \(\rho\) is the population parameter for the ICC, \(\sigma^2\_r\) is the row variability (variability between raters), \(\sigma^2\_c\) is the column variability (variability between faces), and \(\sigma^2\_e\) is the error.

We’re using a two way random model because we expect variability between subjects, but also within (faces have different underlying valence and arousal). Also note that the ‘single measures’ part is referring to the fact that each rating is a single score, not an average of scores.

We’ll use the *ICC* function from the *psych* package to compute the ICCs.

library(psych)

VA\_adult\_icc <- VA\_adult %>%

select(-face) %>%

ICC()

VA\_child\_icc <- VA\_child %>%

select(-face) %>%

ICC()

AR\_adult\_icc <- AR\_adult %>%

select(-face) %>%

ICC()

AR\_child\_icc <- AR\_child %>%

select(-face) %>%

ICC()

Now, we’ll use the *kableExtra* package to generate a table of the results. Note that I’m extracting the 2nd value for the ICC results because it is the ICC2. If we expected no column variability then we might use ICC1.

library(kableExtra)

kable(data.frame(matrix(c(VA\_adult\_icc$results$ICC[2], VA\_child\_icc$results$ICC[2],

AR\_adult\_icc$results$ICC[2], AR\_child\_icc$results$ICC[2]),

nrow = 2, ncol = 2),

row.names = c("Adult", "Child")),

col.names = c("ICC Valence", "ICC Arousal")) %>%

kable\_styling()

|  | **ICC Valence** | **ICC Arousal** |
| --- | --- | --- |
| Adult | 0.806 | 0.194 |
| Child | 0.795 | 0.220 |

Clearly and not surprisingly, the ICCs for arousal (~.20) are much lower than those for valence (~.80). Using Cicchetti’s (1994) guidelines, we would interpret the valence ICCs as indicating excellent agreement and the arousal ICCs as poor agreement. It is also worth noting that adults and children seem equally (un)reliable in their reporting.

**Conclusions**

The findings suggest that we should give pause before attempting to interpret differences between children and adults in their overall ratings of arousal in facial expressions. Such disagreement is actually expected according to dimensional theories of emotion (Russell, 2003) because emotions are not viewed as prototypical things, and there can be wide variability in facial expressions even across similar situations. In other words, there’s no universal facial expression for high arousal (in fact, there’s little reason to believe in universality for any emotional expression).

**References**

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145-172.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing reliability. *Psychological Bulletin, 86*, 420-428.

Vesker, M., Bahn, D., Dege, F., Kauschke, C., & Gudrun, S. (2018). Perceiving arousal and valence in facial expressions: Differences between children and adults. *European Journal of Developmental Psychology, 15*, 411-425.