Design Analysis SSI/FLE

Stefano Dalla Bona

2024-03-27

1. Experimental Procedure

The Sample Size Insensitivity (SSI) task highlights the common difficulty people encounter in recognizing that smaller sample sizes often result in greater variability. The task typically involves the following question:

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

- *The larger hospital*
- *The smaller hospital*
- *About the same (that is, within 5% of each other)*

The correct response is the second alternative. We emphasize that the designed task is brief, typically taking 3/4 minutes to complete.

2. Experiment Design

2.1 Groups

In the original study conducted by Tversky et Kahneman (1974), it was found that only 22% (21 of the 95) of participants were able to accurately recognize that smaller sample sizes can result in increased variability.

```
Response <- c("Smaller (True)","Bigger (False)","Both (False)")
Percent_Or <- c("22%","22%","56%")

data <- data.frame(Response, Percent_Or)

knitr::kable(data)
```

Response	Percent_Or
Smaller (True)	22%
Bigger (False)	22%
Both (False)	56%

The phenomenon in which biases are attenuated through the presentation of tasks in a foreign language is known as the Foreign Language Effect (FLE). According to one theory of the FLE, reasoning in a disfluent foreign language can induce a heightened state of attentiveness, thus promoting the utilization of System 2 analytical reasoning processes (Del Maschio *et al.*, 2022). Our study focuses on unbalanced bilingualism, which Grosjean (2010) defines as a form of bilingualism characterized by an uneven distribution of proficiency across two languages. In this context, one language typically dominates while the other is less used. Although there isn't a universally accepted definition of unbalanced bilingualism, we specifically refer to a type of unbalanced second language that can lead to greater disfluency in its usage.

Thus, we posit that presenting the task in a second language characterized by disfluency (e.g., presenting the task in English for a native Italian speaker who understands the task but experiences difficulty due to limited everyday use of the language) may lead to a reduction in the SSI bias. This, in turn, could enhance individuals' ability to select the correct answer. Conversely, we hypothesize that presenting the task in a fluent second language (e.g., presenting the task in English for a native Italian speaker who learned English from birth and uses it daily) will not lead to a reduction in this cognitive bias.

```
Disfluent <- c("22% + (x)%","22% - (x/y)%","56% - (x/z)%")

Fluent <- c("~22%","~22%","~56%")

data <- data.frame(Response, Percent_Or, Fluent, Disfluent)

knitr::kable(data)
```

Response	Percent_Or	Fluent	Disfluent
Smaller (True)	22%	~22%	22% + (x)%
Bigger (False)	22%	~22%	22% - (x/y)%
Both (False)	56%	~56%	56% - (x/z)%

Here, x represents an arbitrary quantity symbolizing the increase in the disfluent condition. x/y and x/z represent fractions of x, as an increase in one response will correspond to a decrease in other responses.

However, we assert that if we do not find significant results in the difference between the baseline condition and the disfluent language condition, we have no theoretical grounds to suspect that a difference would occur with the fluent second language condition. Therefore, the first experiment will only focus on two conditions: the first language (fluent) and the second language (disfluent). If a statistically relevant result emerges from this comparison, a second experiment can be built upon the findings of the previous one, incorporating the fluent second language condition.

```
knitr::kable(data[,-3])
```

Response	Percent_Or	Disfluent
Smaller (True)	22%	22% + (x)%
Bigger (False)	22%	22% - (x/y)%
Both (False)	56%	56% - (x/z)%

2.2 Checking disfluency

An important aspect of the experiment relies on selecting an appropriate sample with limited proficiency in an unbalanced second language:

A. The recruitment form will target individuals with relatively limited proficiency in their second language. We intend to utilize the Prolific platform for recruitment. The form will explicitly state that we are seeking participants who:

- Are capable of using another language independently.
- Do not possess a language certificate CEFR greater than B2, indicating proficient language skills.
- Are not simultaneous bilinguals (individuals who become bilingual by learning two languages from birth).
- Perceive themselves as not fluent in the second language.
- Do not use the second language every day.

Given the brief nature of the task, we will translate it into several European languages (e.g., English, Spanish,

Italian, German, Portuguese, etc.) to offer a selection of different native languages and disfluent second languages. The primary effect is not expected to vary across different languages. However, we will consider potential variability among languages in a separate analysis. For instance, Italian and Spanish may exhibit greater similarity than Italian and German. We will account for this variability using a random effect parameter in the model (see Section 3), given our pre-selected sample covering not all possible languages.

B. Objective measurements on the task will include:

- Question about possession of a second language diploma.
- Inquiry about the age of acquisition (AOA) of the second language.
- Query regarding the daily usage of the second language.
- Total experiment time, which we anticipate to have a higher median for the experimental group. We expect this because the reading pace is likely to be slower for those reading in a disfluent second language. However, we do not anticipate a statistically significant result due to the brevity of the task.

Cut-offs for these measurements will be determined prior to analysis and preregistered on the Open Science Framework Platform. Only participants who accurately considered the recruitment form required features, providing correct measurements for the objective scales, will be included in the main analyses. Subsequently, the analyses will be conducted again to assess the impact of any outliers that may exist within our sample, albeit expected to be few.

C. Subjective scales commonly utilized in this type of experiments will also be employed:

- Self-assessment of language proficiency.
- Perceived difficulty in reading the task within the experiment, which will help determine if participants genuinely experience disfluency in their second language.

As a secondary analysis (refer to Section 3 for the main analysis), a logistic regression model will be conducted. The model will have the correct responses as the dependent variable and the perceived difficulty in reading and groups as predictors. This analysis aims to evaluate if the difficulty experienced during the task has an impact on the responses.

3. Main hypothesis

Upon initial examination, the data format from the experiment indicate that the Chi-Squared Test of Independence would be well-suited for the experimental design. This is because we have two conditions, each capable of generating three categorical frequencies, thus resulting in a 3x2 matrix. For instance:

```
data$Percent_Or <- c(22,22,56); data$Disfluent <- c(42,12,46)

chisq.test(data[,c(-1,-3)])
```

```
##
##  Pearson's Chi-squared test
##
## data:  data[, c(-1, -3)]
## X-squared = 10.172, df = 2, p-value = 0.006184
```

However, we aim to refine our hypothesis further. Fundamentally, only the proportion of correct responses to the question asked provide meaningful information on bias reduction. The NHST approach in this case is appropriate because our primary objective is to determine whether an effect exists. We assert that presenting the task in a disfluent second language can prompt individuals to increase the number of correct responses. As such, we can formulate two hypotheses:

- *H0: Correct responses proportion (No-FLE) = Correct responses proportion (FLE)*
- *H1: Correct responses proportion (No-FLE) < Correct responses proportion (FLE)*

Where "no-FLE" refers to the control group that performs the task in their native language, where the Foreign Language Effect (FLE) is not expected to be observed, and "FLE" refers to the experimental group

that performs the task in a disfluent second language, where the FLE is expected to be observed.

We can seek to support that the data are unlikely under the null hypothesis by utilizing a One-Tailed Two-Proportion Z test (NHST, see Section 4 for the Design analysis). Additionally, to investigate the differences in proportions, we aim to calculate Bayesian posterior distributions for the two conditions, as detailed in Section 5 where we computed Bayesian priors. Following this, we will assess the estimated proportions of all responses using the Bayesian Chi-Squared method on the 3x2 matrix.The results obtained from this second analysis can subsequently be compared with the estimated proportions of correct answers for the two conditions to validate the robustness of estimated value across different statistical models.

Other more complex (considering the increased number of parameters), secondary models will be constructed and evaluated. These models will incorporate language variability (refer to Section 2.2) as a random variable and the perceived disfluency on the task (refer to Section 2.2) as a potential predictor.

4. Design Analyses (NHST)

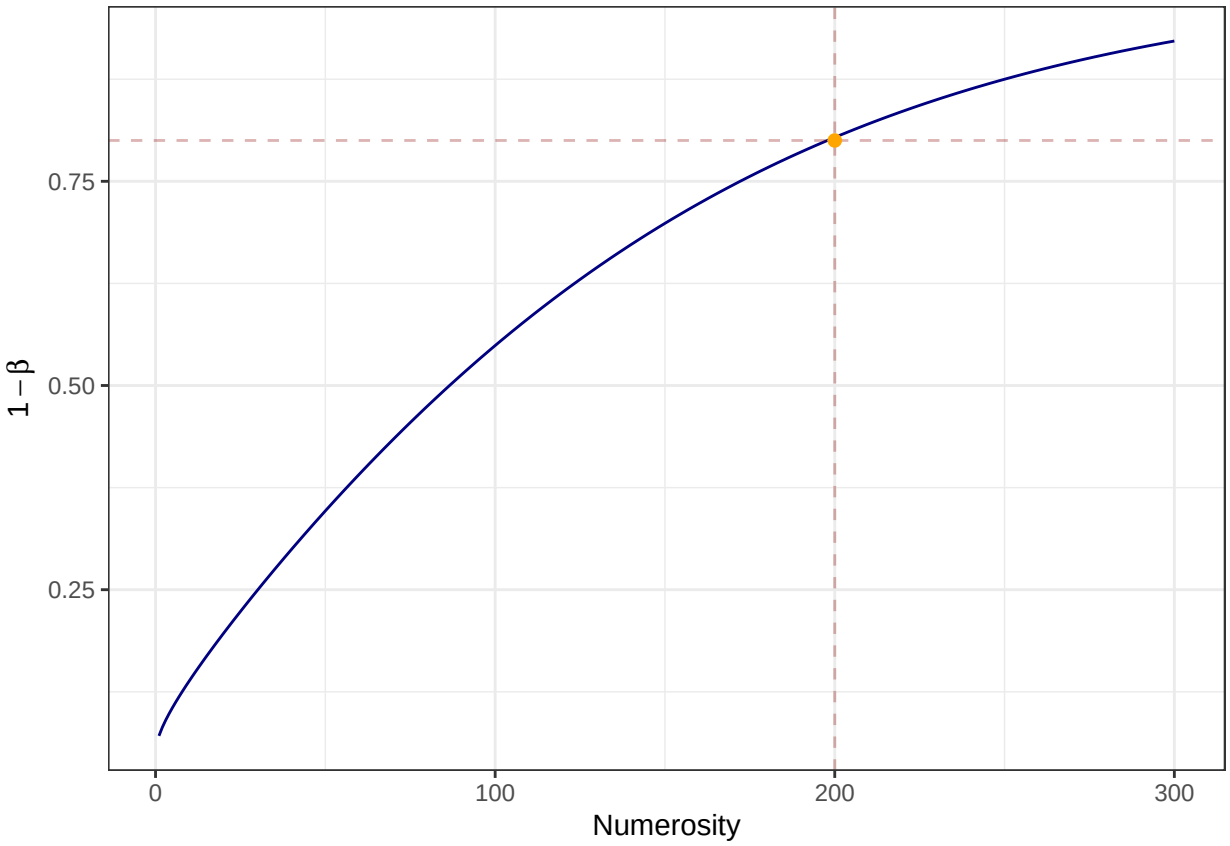
Power is viewed as the complement of *beta*, the false negative rate. The power of the test is the chance to reject the null hypothesis, given the null hypothesis is false. We set the value of *beta* to a standard probability of 0.8 to calculate the necessary sample size for each group. To estimate the Effect Size, we relied on the meta-analyses on the effect of the FLE on biases (typically used in a moral dilemma situation) conducted by *Circi* et al. *(2021)*. Given that the mean effect size expressed in Cohen's d is equal to 0.25 (see Section 5), we are adopting the interpretative label created by Cohen (2013) to specify that we are seeking for a small to medium effect size of Cohen's h, also equal in value to 0.25.

```
library(pwr)
pwr.2p.test(h = .25, sig.level = 0.05, power = .80, alternative = "greater")
```

```
##
##      Difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.25
##               n = 197.8418
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
##
## NOTE: same sample sizes
```

Therefore, we aim to recruit a minimum of 200 participants for each group. We can visualize the power curve as a function of the sample size.

```
x <- c(1:300)
y <- pwr.2p.test(h = .25, sig.level = 0.05, n=x, alternative = "greater")$power
library(ggplot2)
ggplot(mapping=aes(x,y)) + theme_bw()+
  geom_line(color="navy") + xlab("Numerosity") + ylab(expression(1-beta))+
  geom_hline(yintercept = .8, color="darkred", alpha=.3, linetype="dashed")+
  geom_vline(xintercept = 200, color="darkred", alpha=.3,linetype="dashed")+
  geom_point(aes(x=200,y=.8),colour="orange", shape=20, size=3)
```



5. Bayesian Priors

The proportion test within the NHST framework has limitations as it doesn't allow us to incorporate our expert knowledge into the model beforehand. Therefore, we can opt for a Bayesian data analysis approach, specifically using the IBE (Independent Beta Estimation) method. This model assumes that variable x (i.e., the proportion of correct responses in the control group) and variable y (i.e., the proportion of correct responses in the sperimental group) follow independent binomial distributions with success probabilities $\theta(A)$ and $\theta(B)$. These success probabilities are assigned independent $Beta(\alpha, \beta)$ distributions, which encode the relative prior plausibility of values for $\theta(A)$ and $\theta(B)$.

$X \sim Beta(\alpha, \beta)$

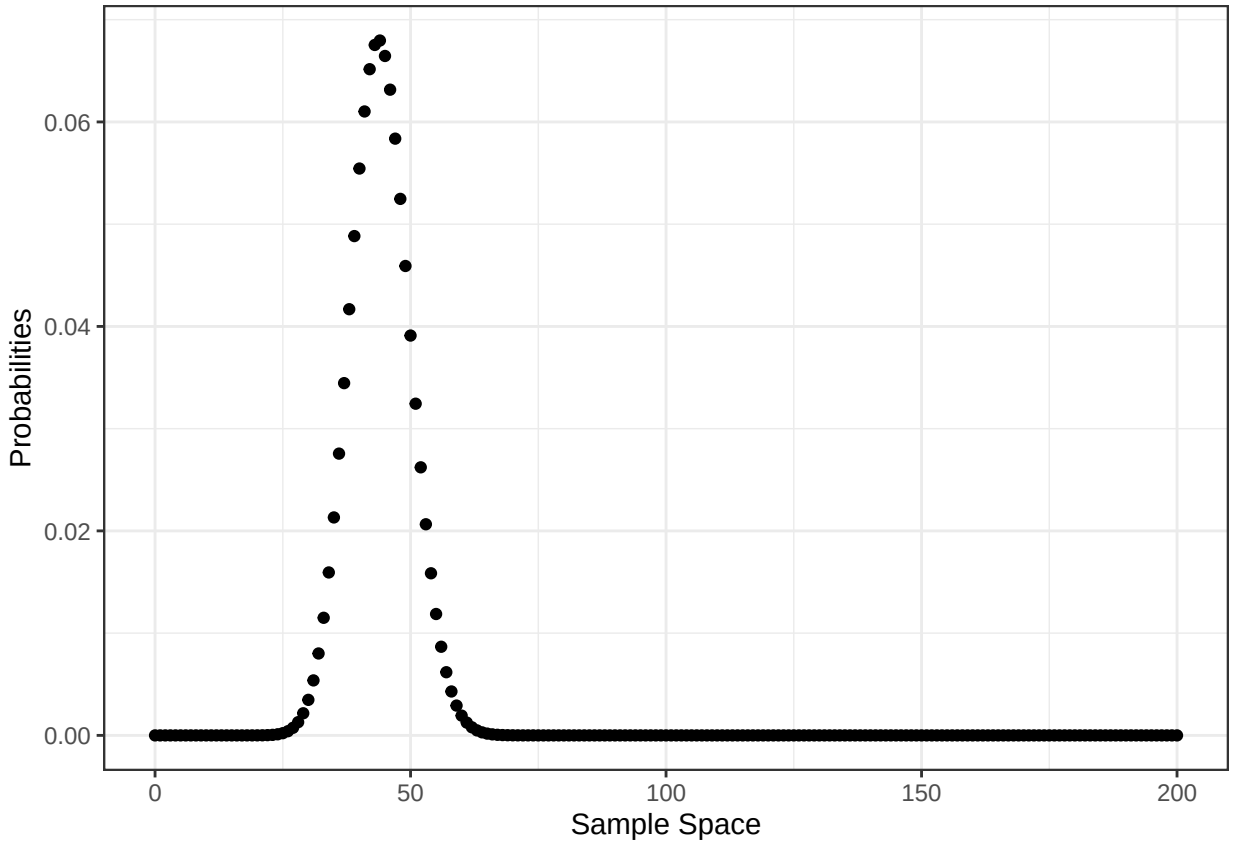
$Y \sim Beta(\alpha, \beta)$

5.1 Control Prior Beta parameters

For the control group, we assumed that we do not have any theoretically valid reason to observe a different result compared to the original study (Tversky et Kahneman, 1974). We then considered the binomial distribution with the π parameter of 0.22 and n equal to our estimated group size from the previous Design analysis (200, see Section 4). We then extracted the mean and variance of this distribution through simulated sampling.

```
x <- 0:200

prob <- dbinom(x, 200, 0.22)


ggplot(mapping=aes(x=x,y=prob))+
```

```
  geom_point()+theme_bw()+labs(x="Sample Space", y="Probabilities")
```




```
c <- replicate(1e4,rbinom(1,200,.22))
c <- c/200
mean(c); var(c)
```

```
## [1] 0.219857
```

```
## [1] 0.0008668062
```

Mean and variance of this distribution can be used to approximate the beta distribution.

```
estBetaParams <- function(mu, var) {
  alpha <- ((1 - mu) / var - 1 / mu) * mu ^ 2
  beta <- alpha * (1 / mu - 1)
  return(params = list(alpha = alpha, beta = beta))
}

estBetaParams(0.22,0.0008)
```

```
## $alpha
## [1] 46.97
##
## $beta
## [1] 166.53
```

5.2 FLE Prior Beta parameters

For the FLE group proportion, we considered metanalysis from *Circi* et al. *(2021)*, that showed that in the moral decision making domain the FLE yields out a small to medium effect size of Hedge's $g = 0.22$, 95% CI [.14, .30]. In the risk-adversion domain the FLE yields out a small to medium effect size of Hedge's $g = .28$, 95% CI [.17, .39]. Hedge's g of all studies can be converted to Cohen's d. We converted every Hedge's g in the metanalysis to a Cohen's d.

```
#Function to convert Hedge's G in Cohen's D
converter <- function(g, N){
  g / (1- (3/((4*N)-9)))
}

a <- c(0.3,0.26,0.22,0.95,0.52,0.64,0.63,0.33,0.4,0.5,0.54,0.3,0.34,0.24,0.38,0.06,0.15,0.34,-0.21,0.12
b <- c(112,80,107,18,328,397,105,152,72,144,211,173,399,202,190,201,223,197,214,242,195,211,209,161,161

data2 <- data.frame(a,b)
cohen <- converter(data2$a,data2$b)
mean(cohen)
```

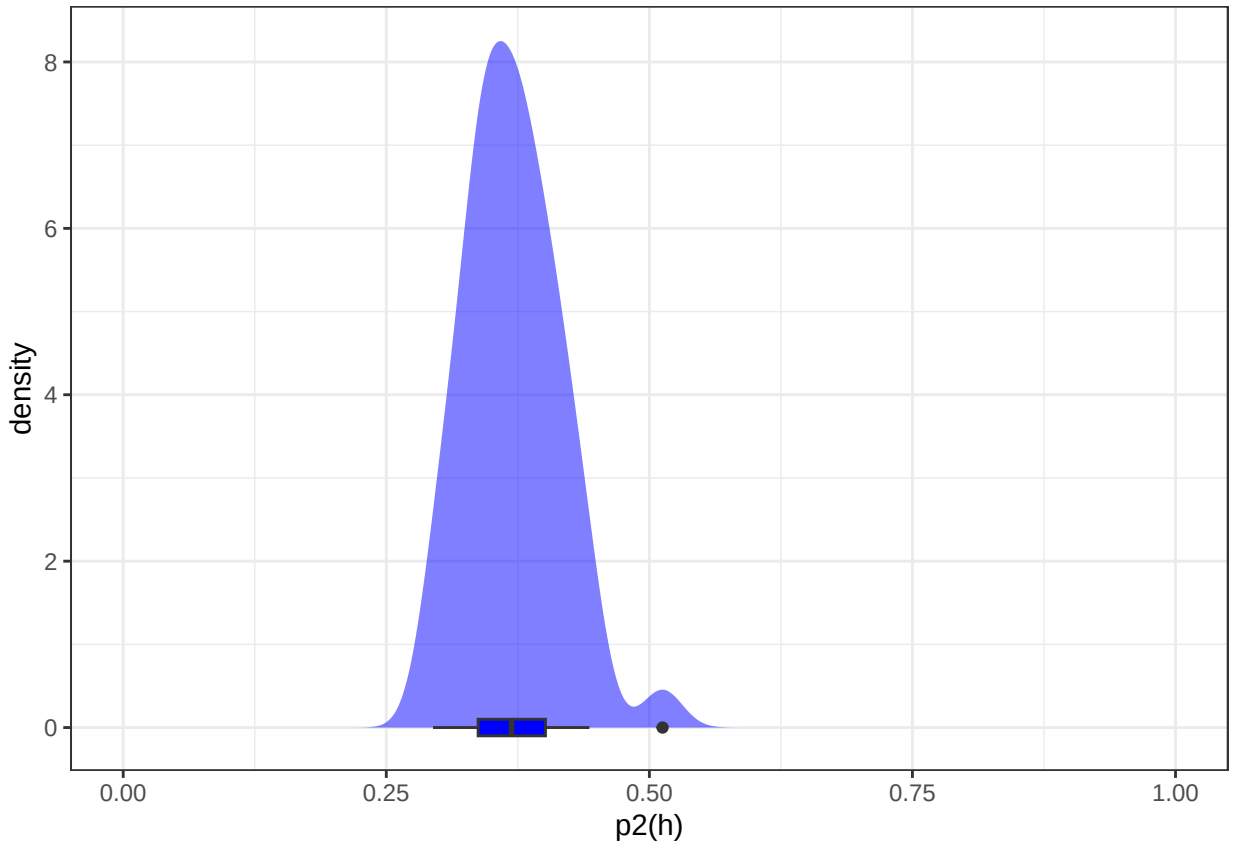
```
## [1] 0.254301
```

Consider that the average Cohen's d approximate 0.25, coherently with our Design Analysis. Cohen's d can then be converted to Odds Ratio (OR).

```
library(effectsize)

odd <- d_to_oddsratio(cohen)
```

An OR is a statistic that quantifies the strength of the association between two events, A and B. The OR is defined as the ratio of the odds of A in the presence of B and the odds of A in the absence of B, or equivalently (due to symmetry), the ratio of the odds of B in the presence of A and the odds of B in the absence of A. We can convert OR in proportions, given the estimed proportion of 0.22 for the control group. From the effect size we can perform the inverse formula to compute the density of expected proportions from all studies considered and extract descriptive statistics.

```
h <- (odd*0.22) / (1+(odd*.22)-.22)

p2 <- function(x){sin((x+(2*asin(0.22)))/2)}

ggplot(mapping = aes(p2(h)))+
  geom_density(fill="blue", color=NA, alpha=.5)+theme_bw()+
  geom_boxplot(width=.2, fill="blue")+xlim(c(0,1))
```




```
pastecs::stat.desc(p2(h))
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 47.000000000  0.000000000  0.000000000  0.294449924  0.512538933  0.218089009
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 17.417822582  0.369019725  0.370591970  0.006523548  0.013131221  0.002000164
##      std.dev     coef.var
##  0.044723191  0.120680411
```

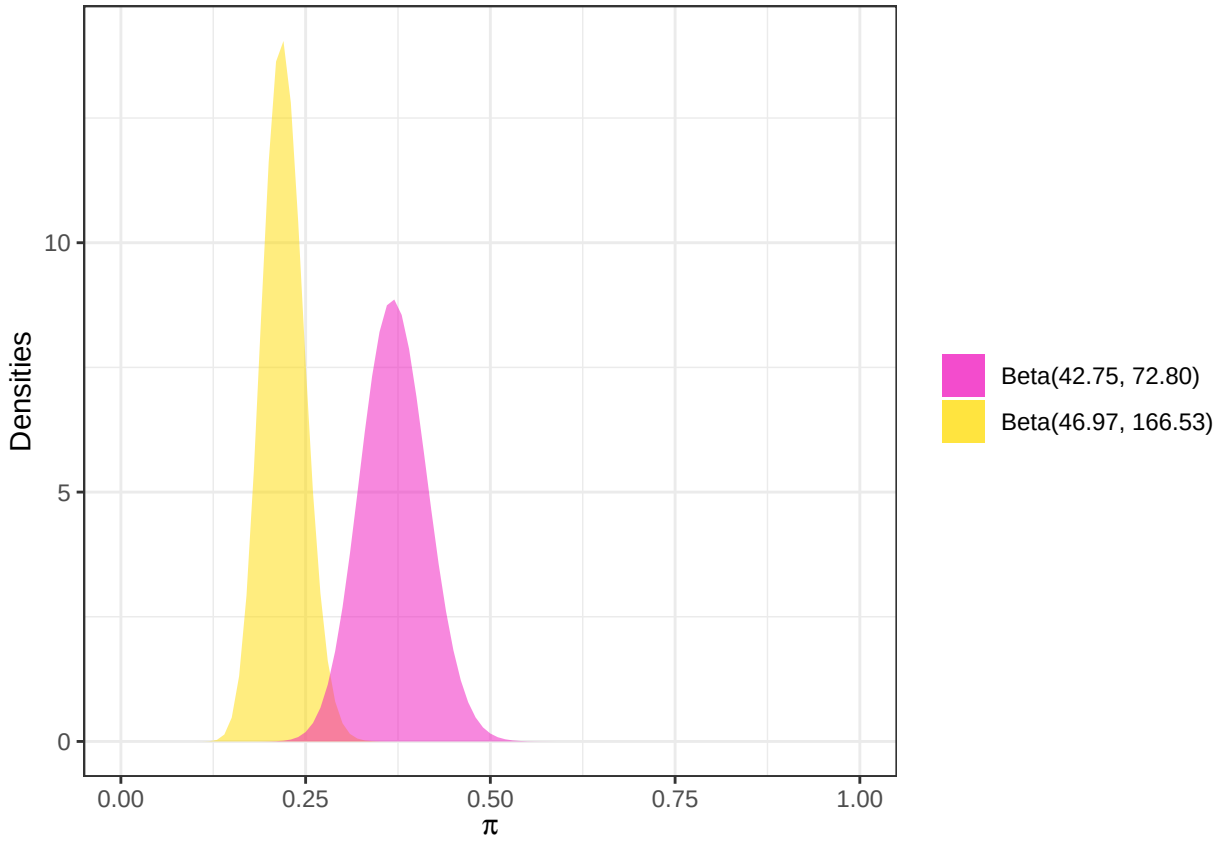
Thus, since that for the FLE group we have to employ a beta distribution, we can build the prior beta distribution as having the same central and dispersion statistics (M = 0.31, Var = 0.009) of this calculated distribution from the metanalysis.

```
estBetaParams <- function(mu, var) {
  alpha <- ((1 - mu) / var - 1 / mu) * mu ^ 2
  beta <- alpha * (1 / mu - 1)
  return(params = list(alpha = alpha, beta = beta))
}

estBetaParams(0.37,0.002)
```

```
## $alpha
## [1] 42.7535
##
## $beta
## [1] 72.7965
```

5.3 Showing Beta Priors

```
ggplot() +
  stat_function(fun = ~dbeta(., 46.97,166.53), geom = "area",
                aes(fill = "Beta(46.97, 166.53)"), alpha = 0.5) +
  stat_function(fun = ~dbeta(., 42.75, 72.80), geom = "area",
                aes(fill = "Beta(42.75, 72.80)"), alpha = 0.5) +
  scale_fill_manual(values = c("Beta(46.97, 166.53)" = "#FFDC00",
                               "Beta(42.75, 72.80)" = "#F012BE")) +
  labs(x = expression(pi), y = "Densities", fill = NULL) +
  theme(axis.text.y = element_blank(),
        legend.position = "top")+
  theme_bw()
```



6. References

Circi, R., Gatti, D., Russo, V. et Vecchi, T. (2021). The foreign language effect on decision-making: A meta-analysis. *Psychonomic Bulletin & Review*, 1-11.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Del Maschio, N., Crespi, F., Peressotti, F., Abutalebi, J. et Sulpizio, S. (2022). Decision-making depends on language: A meta-analysis of the Foreign Language Effect. *Bilingualism: Language and Cognition*, *25*(4), 617-630.

Grosjean, F. (2010). *Bilingual: Life and reality*. Harvard university press.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org

Tversky, A. et Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124-1131.