Measurement robustness for frequency of vowels

Introduction

Everett (2017) uses the frequency of vowels in the basic vocabulary (from the ASJP). We can test the measurement robustness of this by looking at a wider range of words. The database of words from Slonimska & Roberts (2017) was used (A collection of data from the Intercontinental Dictionary Series data, the World Loanword Database and Spraakbanken). It consists of 999 concepts in 226 languages. These were linked to the ASJP estimates for number of vowels.

Test

Load libraries:

```
library(lme4)
library(ggplot2)
library(lattice)
library(sjPlot)
```

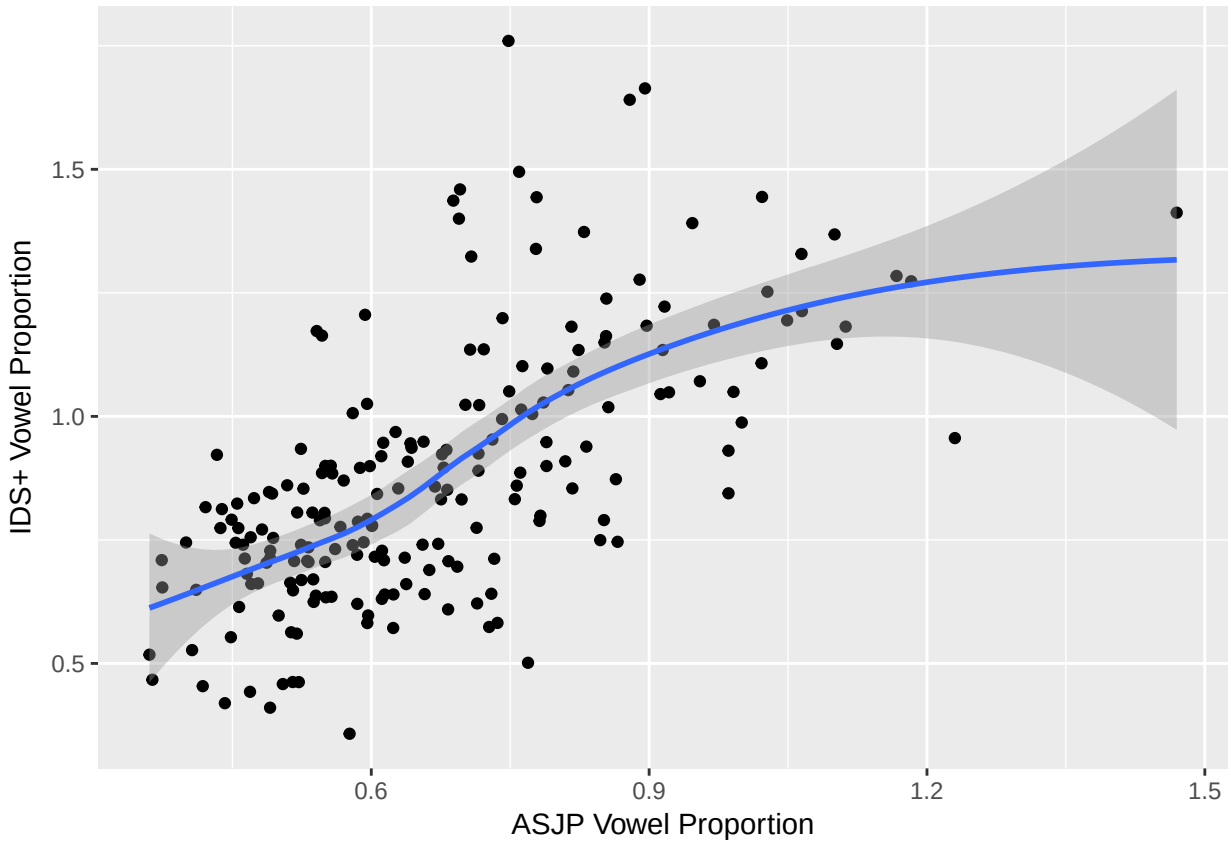
Load data:

```
d = read.csv("../data/ASJP_and_SlonimskaRoberts_VowelProportions.csv", stringsAsFactors = F)
d = d[complete.cases(d[,c("S.VProp","asjp.VProp")]),]
```

Plot raw data

```
ggplot(d[!is.na(d$asjp.VProp),], aes(x =asjp.VProp, y = S.VProp)) +
         xlab("ASJP Vowel Proportion") +
         ylab("IDS+ Vowel Proportion") +
         geom_point() +
         geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



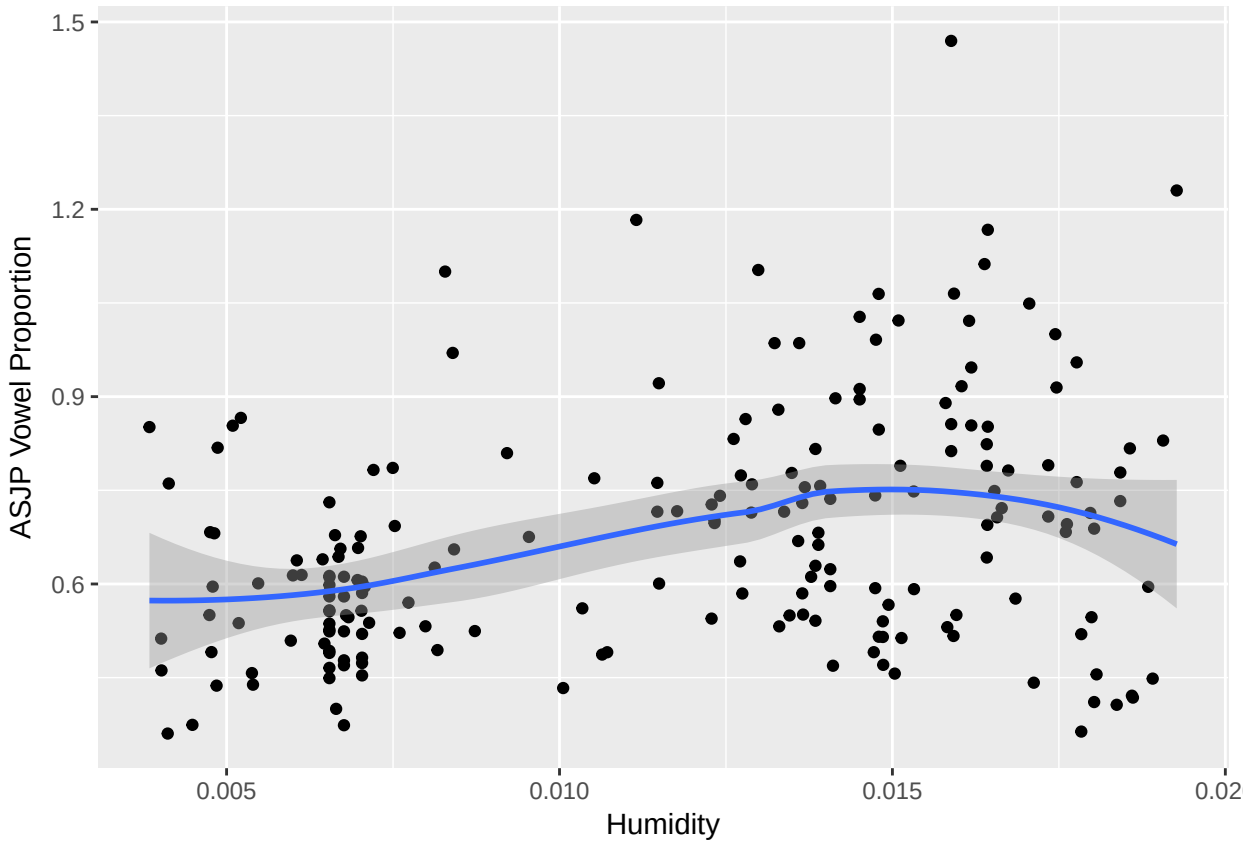
Correlation between ASJP and IDS+ frequencies.

```
cor.test(d$asjp.VProp, d$S.VProp)
```

```
##
##  Pearson's product-moment correlation
##
## data:  d$asjp.VProp and d$S.VProp
## t = 11.966, df = 199, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5584831 0.7207496
## sample estimates:
##       cor
## 0.6468798
```

```
ggplot(d[!is.na(d$asjp.VProp),], aes(x =specH.mean, y = asjp.VProp)) +
  xlab("Humidity") +
  ylab("ASJP Vowel Proportion") +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## Warning: Removed 4 rows containing non-finite values (stat_smooth).
## Warning: Removed 4 rows containing missing values (geom_point).
```



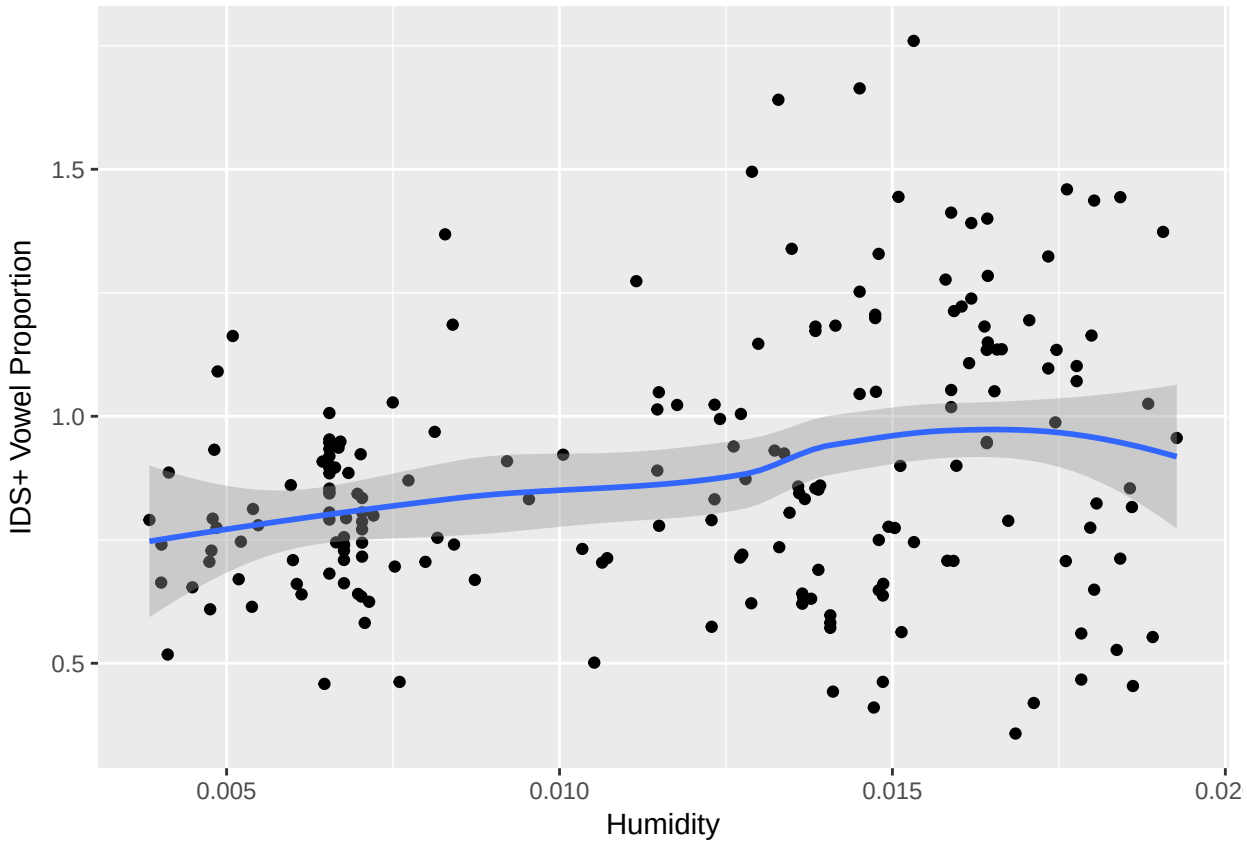
Plot the relationship between humidity and IDS+ esimate. Although note that the relationship between humidity and proportion of vowels does not hold in the sub-sample even when using the ASJP data. This is probably due to lack of power rather than a failure to replicate the original result (which used 10 times more data).

```
ggplot(d, aes(x =specH.mean, y = S.VProp)) +
  xlab("Humidity") +
  ylab("IDS+ Vowel Proportion") +
  geom_point() +
  geom_smooth()
```

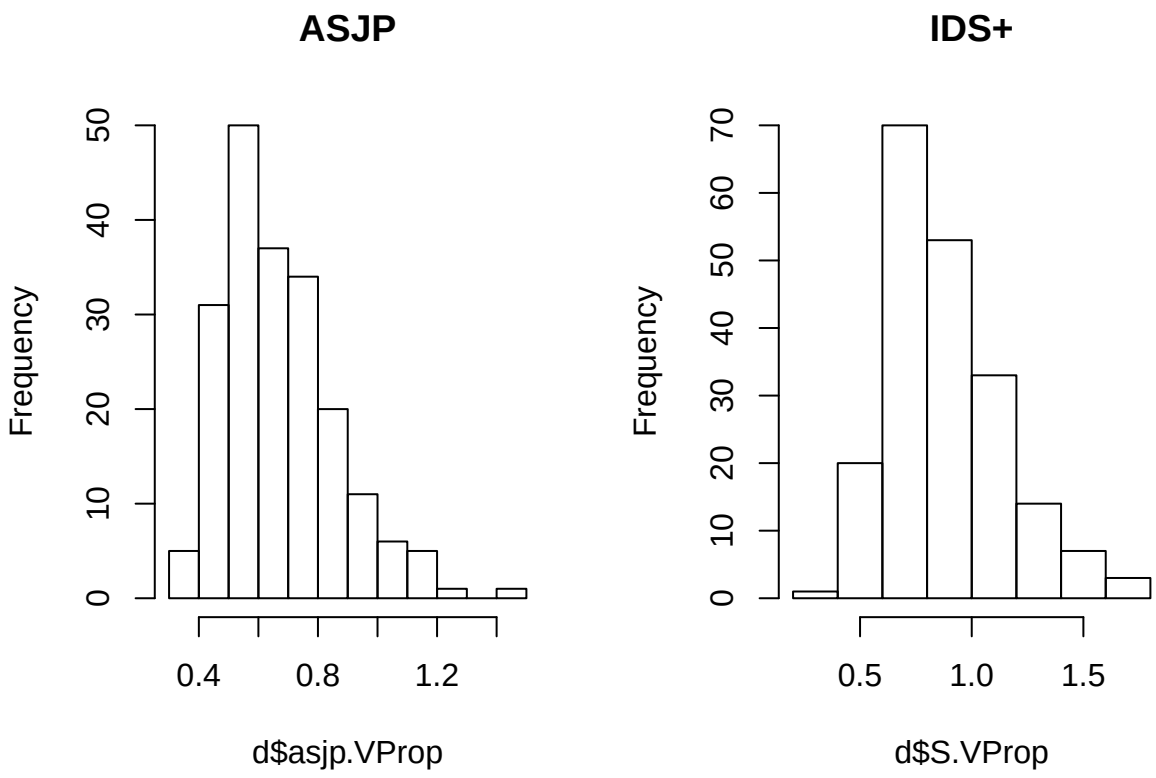
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 4 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```




```
par(mfrow = c(1,2))
hist(d$asjp.VProp, main="ASJP")
hist(d$S.VProp,main="IDS+")
```




```
par(mfrow=c(1,1))

d$asjp.VProp.scale = scale(d$asjp.VProp)
d$S.VProp.scale = scale(d$S.VProp)


m0 = lmer(asjp.VProp.scale~ 1 +
            (1  | family) +
            (1  | autotyp.area),
          data = d[!is.na(d$specH.mean),])

m1 = lmer(asjp.VProp.scale~ 1 +
            specH.mean +
            (1  | family) +
            (1  | autotyp.area),
          data = d[!is.na(d$specH.mean),])

anova(m0,m1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d[!is.na(d$specH.mean), ]
## Models:
## m0: asjp.VProp.scale ~ 1 + (1 | family) + (1 | autotyp.area)
## m1: asjp.VProp.scale ~ 1 + specH.mean + (1 | family) + (1 | autotyp.area)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m0  4 456.72 469.85 -224.36   448.72
## m1  5 458.00 474.42 -224.00   448.00 0.7114      1      0.399
```

```
m0 = lmer(S.VProp.scale~ 1 +
          (1 + specH.mean | family) +
```

```
          (1 + specH.mean | autotyp.area),
            data = d,
          control = lmerControl(optimizer = 'bobyqa',
                                optCtrl = list(maxfun=50000)))
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
m1 = lmer(S.VProp.scale~ 1 +
            specH.mean +
            (1 + specH.mean | family) +
            (1 + specH.mean | autotyp.area),
          data = d,
          control = lmerControl(optimizer = 'bobyqa',
                                optCtrl = list(maxfun=50000)))

anova(m0,m1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
## Models:
## m0: S.VProp.scale ~ 1 + (1 + specH.mean | family) + (1 + specH.mean |
## m0:     autotyp.area)
## m1: S.VProp.scale ~ 1 + specH.mean + (1 + specH.mean | family) +
## m1:     (1 + specH.mean | autotyp.area)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m0  8 426.41 452.68 -205.21   410.41
## m1  9 425.77 455.32 -203.89   407.77 2.6401      1     0.1042
```

Test of bias

Are the differences between the estimates biased? The exact estimate difference doesn't matter, only the relative relationship. So we use the residuals from a linear model as an indication of how different the estimates are. When we come to adding a fixed effect for humidity, this also avoids haivng to test a relationship between the proporiton of vowels and humidity, which is part of the main hypothesis.

```
linear.model = lm(S.VProp ~ asjp.VProp, data=d)

d$Vowel.diff = resid(linear.model)

m.full = lmer(Vowel.diff ~
                (1| family) +
                (1| autotyp.area),
              data = d)

m.noFamily = lmer(Vowel.diff ~
                (1| autotyp.area),
              data = d)

m.noArea = lmer(Vowel.diff ~
                (1| family),
              data = d)

anova(m.full,m.noFamily)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
## Models:
## m.noFamily: Vowel.diff ~ (1 | autotyp.area)
## m.full: Vowel.diff ~ (1 | family) + (1 | autotyp.area)
##            Df     AIC     BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m.noFamily  3 -115.85 -105.94 60.924  -121.85
## m.full      4 -142.83 -129.62 75.414  -150.83 28.98      1  7.313e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m.full,m.noArea)
```

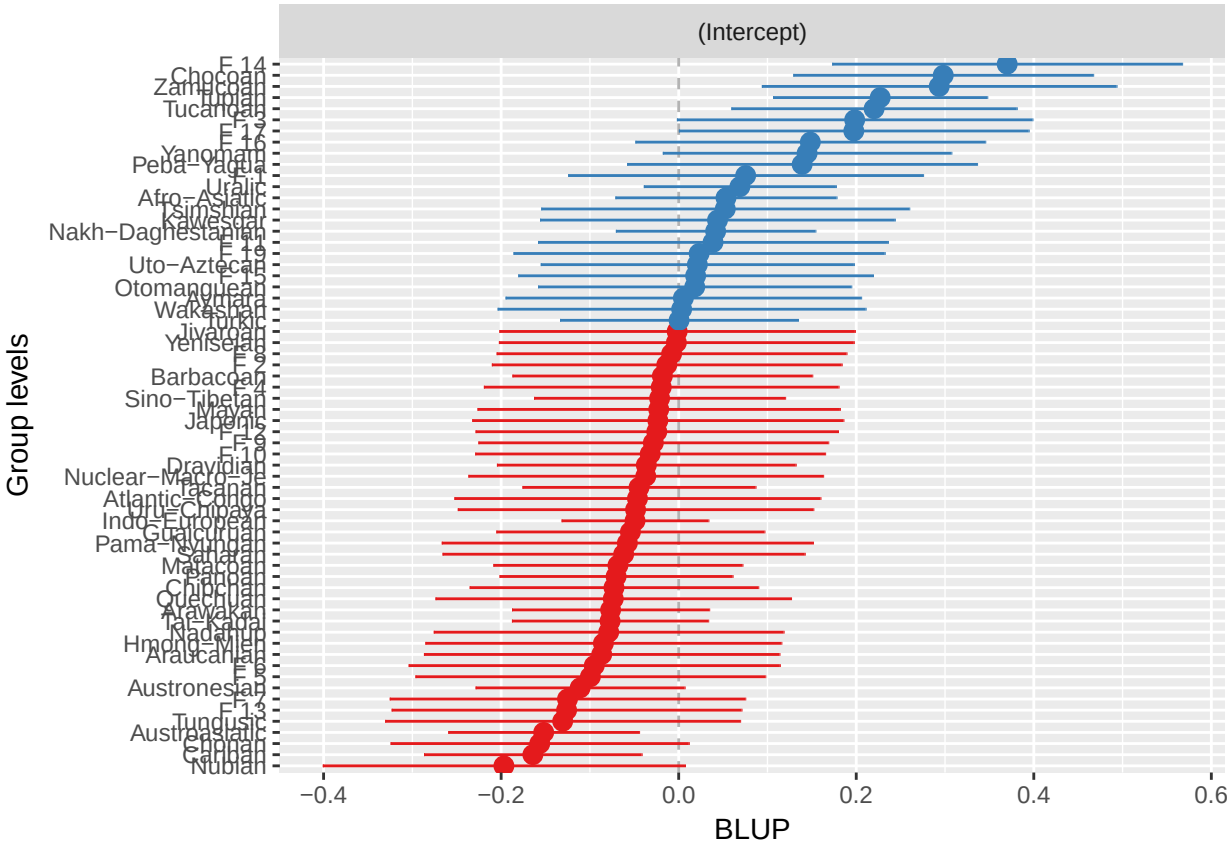
```
## refitting model(s) with ML (instead of REML)
```

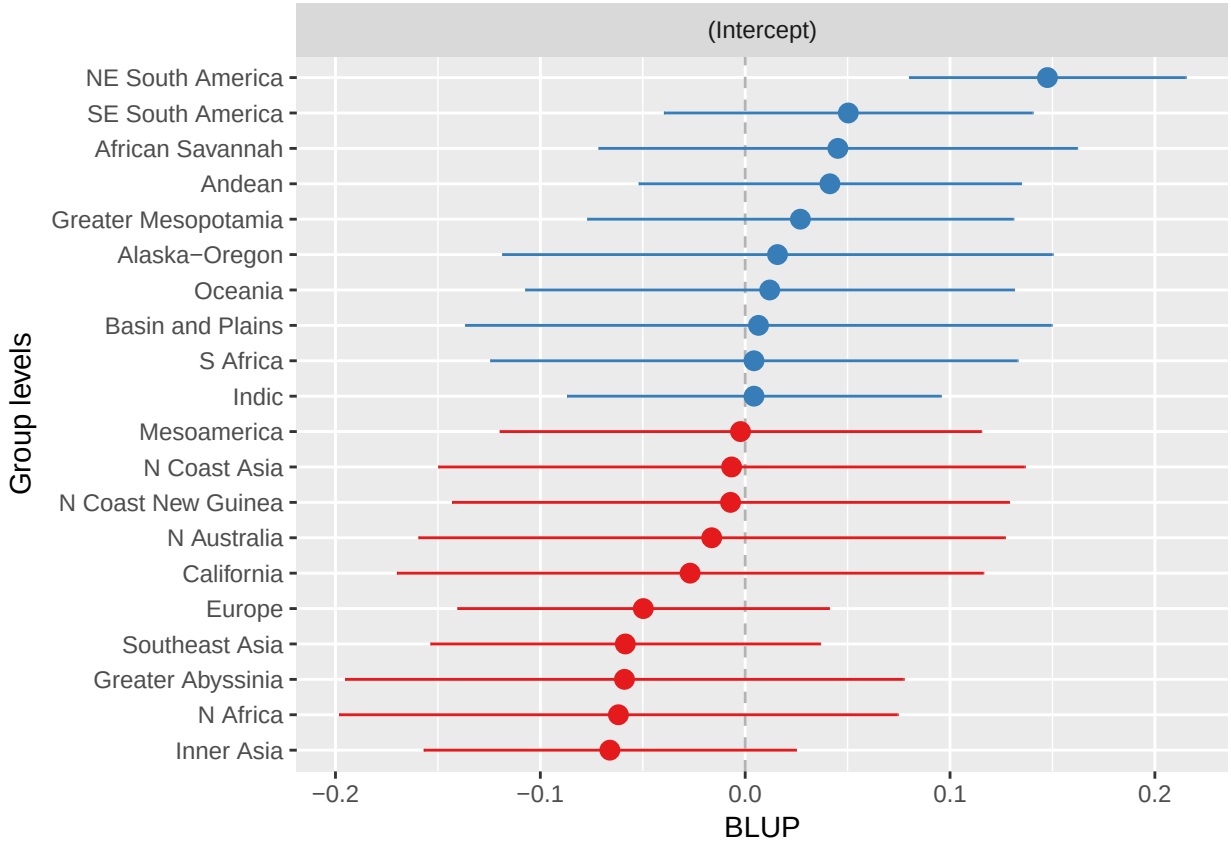
```
## Data: d
## Models:
## m.noArea: Vowel.diff ~ (1 | family)
## m.full: Vowel.diff ~ (1 | family) + (1 | autotyp.area)
##          Df     AIC     BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## m.noArea  3 -135.63 -125.72 70.815  -141.63
## m.full    4 -142.83 -129.62 75.414  -150.83 9.1997      1   0.002421 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding family or area as a random effect significantly improves the fit of the model. The ASJP measures are more weakly related to the IDS+ measures in some families and some areas (the ones with lower values in this table):

```
sjp.lmer(m.full, 're', show.values = F, sort.est = "(Intercept)")
```

```
## Plotting random effects...
## Plotting random effects...
```






```
d$specH.mean = scale(d$specH.mean)
m.full2 = lmer(Vowel.diff ~
                 (1| family) +
                 (1| autotyp.area),
               data = d[!is.na(d$specH.mean),])
m.familyHslope = lmer(Vowel.diff ~
                 (1 + specH.mean| family) +
                 (1| autotyp.area),
               data = d[!is.na(d$specH.mean),])
m.areaHslope = lmer(Vowel.diff ~
                 (1 + specH.mean| family) +
                 (1 + specH.mean| autotyp.area),
               data = d[!is.na(d$specH.mean),])

anova(m.full2, m.familyHslope, m.areaHslope)
```

```
## refitting model(s) with ML (instead of REML)
```

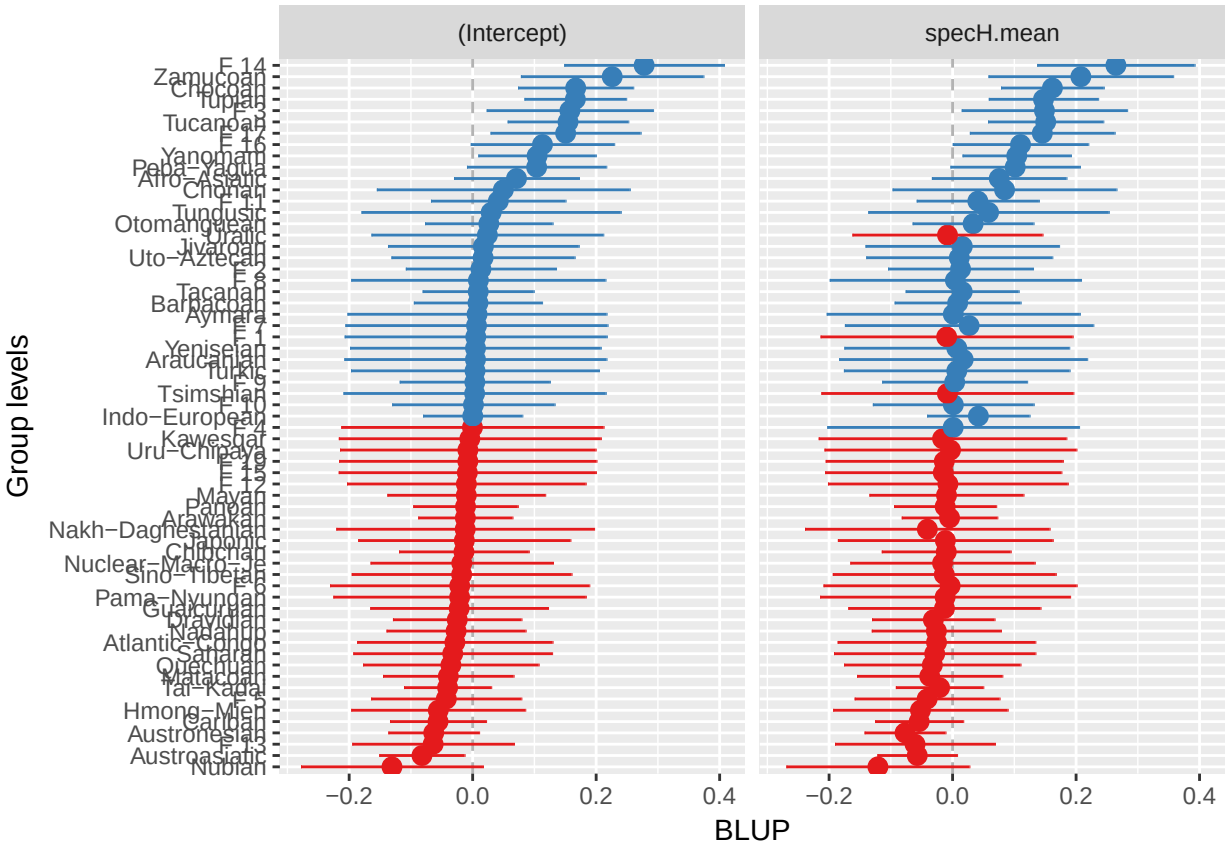
```
## Data: d[!is.na(d$specH.mean), ]
## Models:
## m.full2: Vowel.diff ~ (1 | family) + (1 | autotyp.area)
## m.familyHslope: Vowel.diff ~ (1 + specH.mean | family) + (1 | autotyp.area)
## m.areaHslope: Vowel.diff ~ (1 + specH.mean | family) + (1 + specH.mean | autotyp.area)
##                Df     AIC     BIC logLik deviance   Chisq Chi Df
## m.full2         4 -136.42 -123.29 72.212  -144.42
## m.familyHslope  6 -143.53 -123.83 77.766  -155.53 11.1092      2
## m.areaHslope    8 -143.57 -117.30 79.784  -159.57  4.0354      2
##                Pr(>Chisq)
```

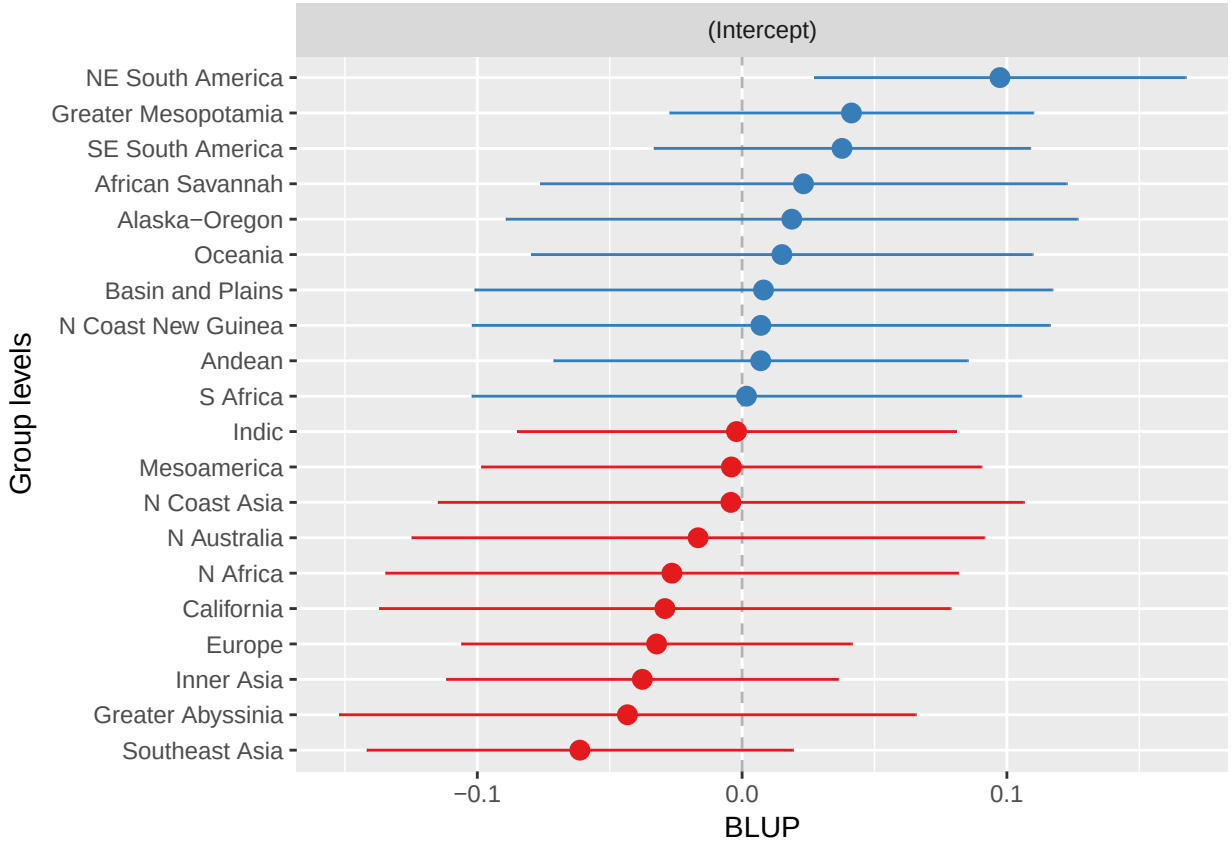
```
## m.full2
## m.familyHslope    0.00387 **
## m.areaHslope      0.13296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding a random slope for humidity by language family improves the fit of the model. This suggests that there is a bias according to humidity: In some families there is

```
sjp.lmer(m.familyHslope, 're', show.values = F, sort.est = "(Intercept)")
```

```
## Plotting random effects...
## Plotting random effects...
```





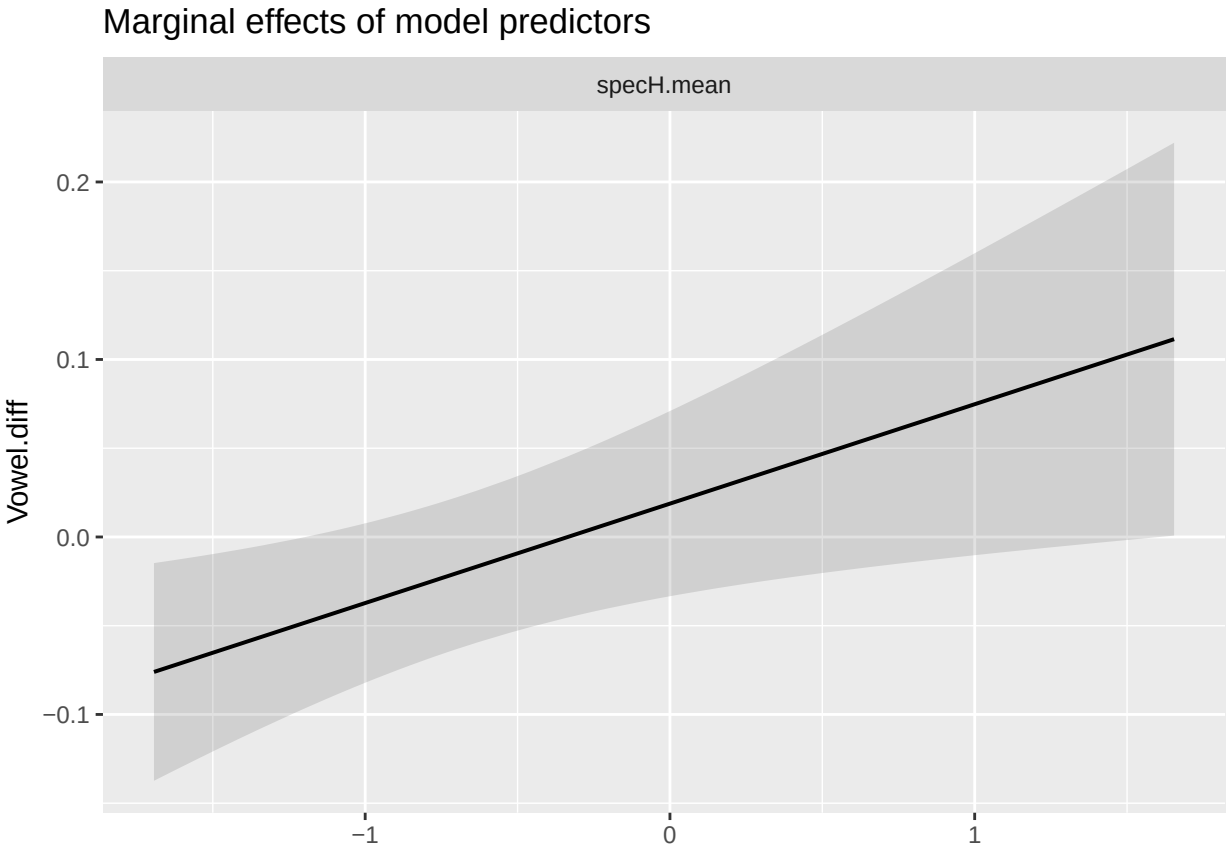
Test main effect of humidity

```
m.areaHfix = lmer(Vowel.diff ~
                    specH.mean +
                (1 + specH.mean| family) +
                (1 + specH.mean| autotyp.area),
              data = d[!is.na(d$specH.mean),])
anova(m.areaHslope,m.areaHfix)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d[!is.na(d$specH.mean), ]
## Models:
## m.areaHslope: Vowel.diff ~ (1 + specH.mean | family) + (1 + specH.mean | autotyp.area)
## m.areaHfix: Vowel.diff ~ specH.mean + (1 + specH.mean | family) + (1 + specH.mean |
## m.areaHfix:     autotyp.area)
##              Df     AIC     BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## m.areaHslope  8 -143.57 -117.30 79.784  -159.57
## m.areaHfix    9 -145.07 -115.53 81.537  -163.07 3.5062      1    0.06114 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sjp.lmer(m.areaHfix,'eff', show.ci = T)
```



The result is marginal. This suggests that there is a weak bias for the differences to be larger in more humid regions.