# Word Relatedness Assignment 1

*Fleur Petit(5583837), Jorrit Jorritsma(3845567), Debby Lam (4298772)*

*22 February 2019*

## Columns and classes of the data

Column "source" has been added to indicate the source-file of the wordpair. I.e. either the file with dissimilar or with similar words.

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   level = col_character(),
##   study_type = col_character(),
##   profession = col_character(),
##   start_time = col_datetime(format = ""),
##   end_time = col_datetime(format = "")
## )

## See spec(...) for full column specifications.

## Parsed with column specification:
## cols(
##   wordpairs = col_character(),
##   similarity = col_character()
## )

## Joining, by = "word_pair"
```

## Mean, standard deviation, lower and upper 95% confidence interval, median of the word pairs

Participants were asked to rate the relatedness of word pairs. More specifically, we introduced the experiment as follows:

"In this experiment you will have to indicate how closely related pairs of words are. For example `flying` and `airplane` are closely related, while `pear` and `acceptance` are not. At the start of the experiment you will furthermore be asked to indicate your studies of profession."

With each word pair we asked:

"Please rate how closely the words are related. 0 means that they are not related at all. 10 means that they are very closely related."

In the below table one can see the descriptive statistics of relatedness ratings of each word pair.

| word_pair | mean | sd | lower | upper | median | source |
|---|---|---|---|---|---|---|
| archive_bird | 1.1666667 | 1.4652846 | 0.4897522 | 1.843581 | 1.0 | dissimilar |
| defeating_discipline | 2.8888889 | 2.2462751 | 1.8511818 | 3.926596 | 2.0 | dissimilar |
| fertility_crane | 3.6111111 | 3.0513042 | 2.2015064 | 5.020716 | 2.0 | dissimilar |
| fertility_hotel | 1.1111111 | 1.0786096 | 0.6128280 | 1.609394 | 1.0 | dissimilar |
| governor_jazz | 0.6111111 | 1.2432826 | 0.0367544 | 1.185468 | 0.0 | dissimilar |
| motto_Jackson | 1.0000000 | 1.3284223 | 0.3863115 | 1.613689 | 0.5 | dissimilar |
| number_equipment | 2.7222222 | 2.1089600 | 1.7479503 | 3.696494 | 2.0 | dissimilar |
| phone_artifact | 3.0555556 | 2.5315009 | 1.8860833 | 4.225028 | 2.5 | dissimilar |

| word_pair | mean | sd | lower | upper | median | source |
|---|---|---|---|---|---|---|
| possession_popcorn | 1.2222222 | 1.3956047 | 0.5774976 | 1.866947 | 1.0 | dissimilar |
| professor_currency | 0.7777778 | 0.9428090 | 0.3422302 | 1.213325 | 0.5 | dissimilar |
| round_Jerusalem | 1.0000000 | 1.2366939 | 0.4286871 | 1.571313 | 0.5 | dissimilar |
| size_environment | 3.3333333 | 2.9305691 | 1.9795044 | 4.687162 | 3.0 | dissimilar |
| travel_soap | 2.5000000 | 2.4793263 | 1.3546307 | 3.645369 | 2.0 | dissimilar |
| video_problem | 1.6666667 | 2.3008950 | 0.6037270 | 2.729606 | 1.0 | dissimilar |
| war_journal | 4.8333333 | 2.6623740 | 3.6034019 | 6.063265 | 5.0 | dissimilar |
| artifact_object | 7.7222222 | 1.9645552 | 6.8146607 | 8.629784 | 8.0 | similar |
| collection_memorabilia | 6.8333333 | 2.0364329 | 5.8925666 | 7.774100 | 7.0 | similar |
| drug_alcohol | 8.3888889 | 1.3779306 | 7.7523291 | 9.025449 | 8.5 | similar |
| flight_airport | 8.7222222 | 1.1785113 | 8.1777878 | 9.266657 | 9.0 | similar |
| gem_jewel | 9.3888889 | 0.6978023 | 9.0665266 | 9.711251 | 9.5 | similar |
| gin_alcohol | 8.8333333 | 1.0981267 | 8.3260340 | 9.340633 | 9.0 | similar |
| investigation_operation | 4.3888889 | 2.2265481 | 3.3602951 | 5.417483 | 4.0 | similar |
| minister_bishop | 5.4444444 | 2.8330450 | 4.1356685 | 6.753220 | 6.0 | similar |
| ministry_government | 8.5555556 | 1.4234268 | 7.8979780 | 9.213133 | 9.0 | similar |
| music_jazz | 8.6111111 | 1.3779306 | 7.9745513 | 9.247671 | 9.0 | similar |
| practice_law | 6.6666667 | 1.4950900 | 5.9759830 | 7.357350 | 7.0 | similar |
| radio_music | 8.6666667 | 1.1375929 | 8.1411352 | 9.192198 | 9.0 | similar |
| science_scientist | 8.8333333 | 1.6179144 | 8.0859087 | 9.580758 | 9.0 | similar |
| troops_personnel | 4.7777778 | 2.7771241 | 3.4948355 | 6.060720 | 6.0 | similar |
| year_century | 8.6666667 | 1.3284223 | 8.0529781 | 9.280355 | 9.0 | similar |

The relatedness ratings are very different for word pairs from the "similar" dataset and word pairs from the "dissimilar" dataset (see the table below). This may indicate that word relatedness and word similarity are rated according to similar standards.

| source | mean | sd | lower | upper | median |
|---|---|---|---|---|---|
| dissimilar | 2.100000 | 2.347426 | 1.015564 | 3.184436 | 1 |
| similar | 7.633333 | 2.309428 | 6.566452 | 8.700215 | 8 |

**Density plots of relatedness rateings per word pair with mean and 95% confidence interval**

Density plots of relatedness ratings with mean and 95% confidence interval
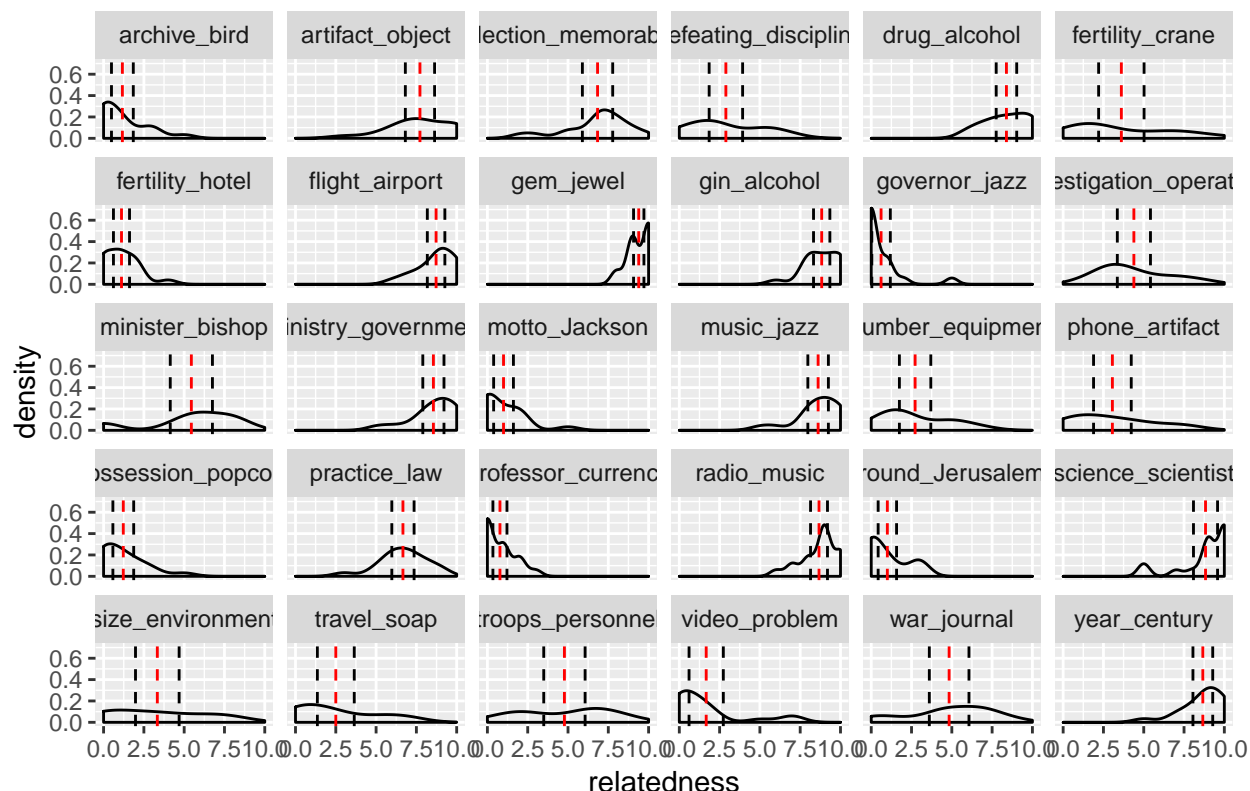


## Pairs with a relatively large spread

A sd > 2.5 has been chosen arbitrarily to indicate a relatively large spread.

| word_pair | sd |
|---|---|
| phone_artifact | 2.531501 |
| war_journal | 2.662374 |
| troops_personnel | 2.777124 |
| minister_bishop | 2.833045 |
| size_environment | 2.930569 |
| fertility_crane | 3.051304 |

## Relatedness rating frequencies per participant

It can be noticed that there are differences in how participants rate the word pairs. Some participants go for extreme ratings more often, while others rate more gradually. The different groups do not seem to give an explanation for this in the graph below.
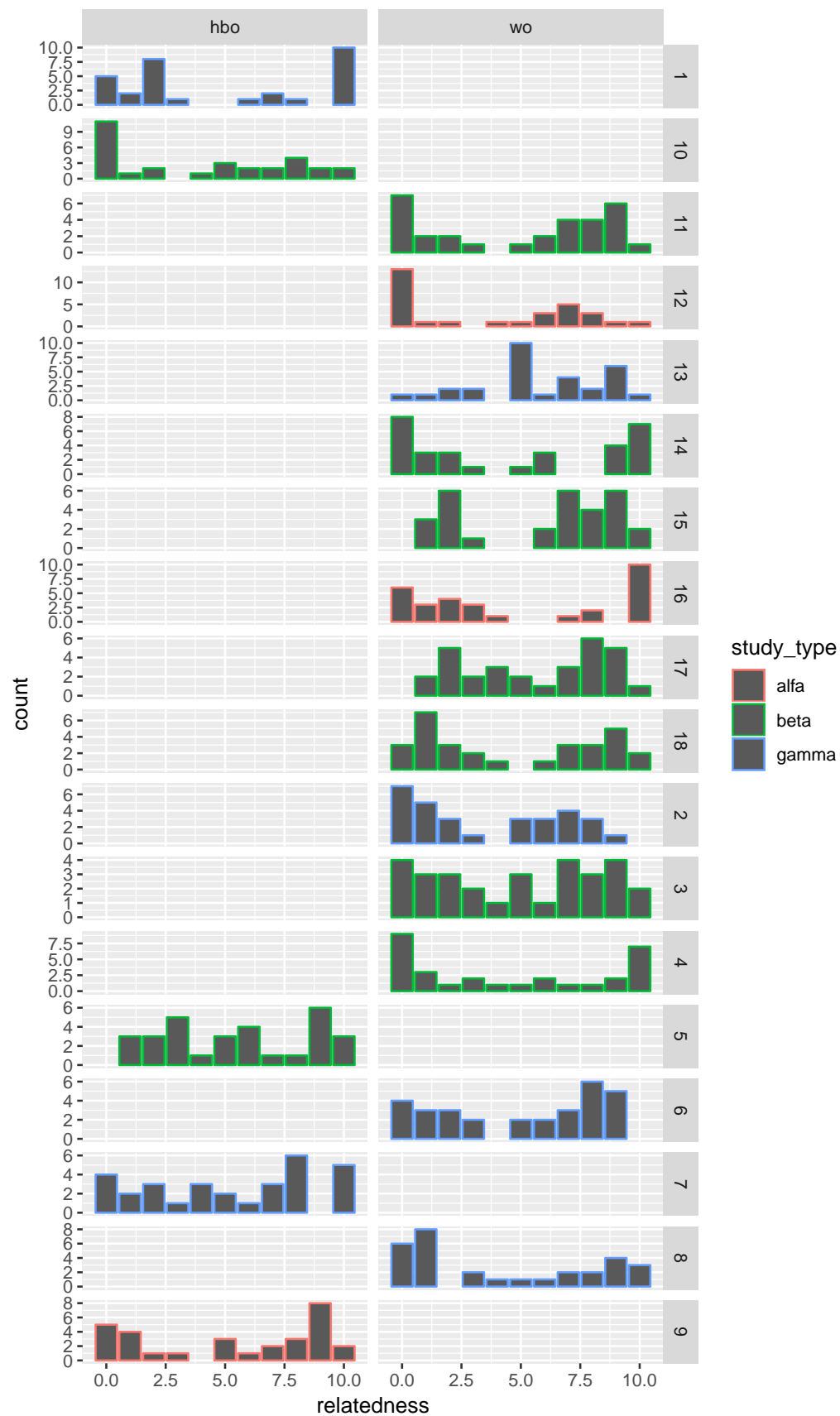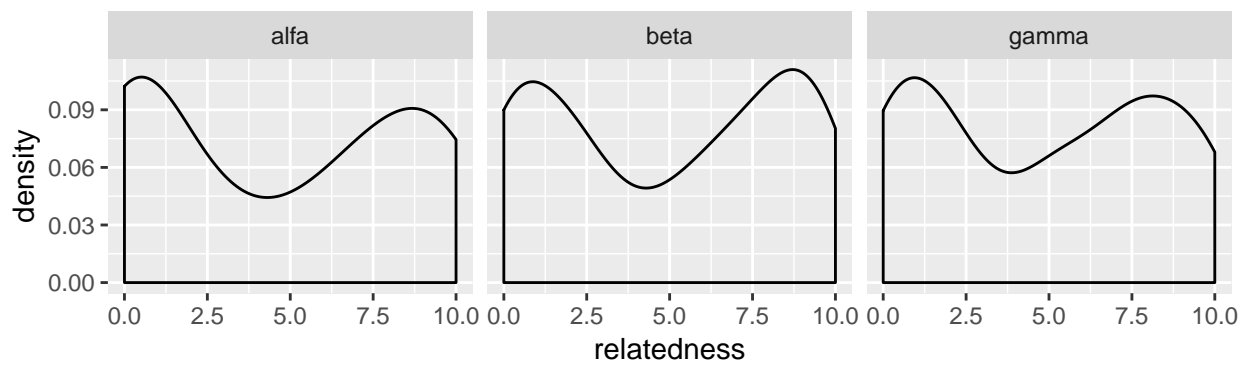
## Distributions of participants among different groups
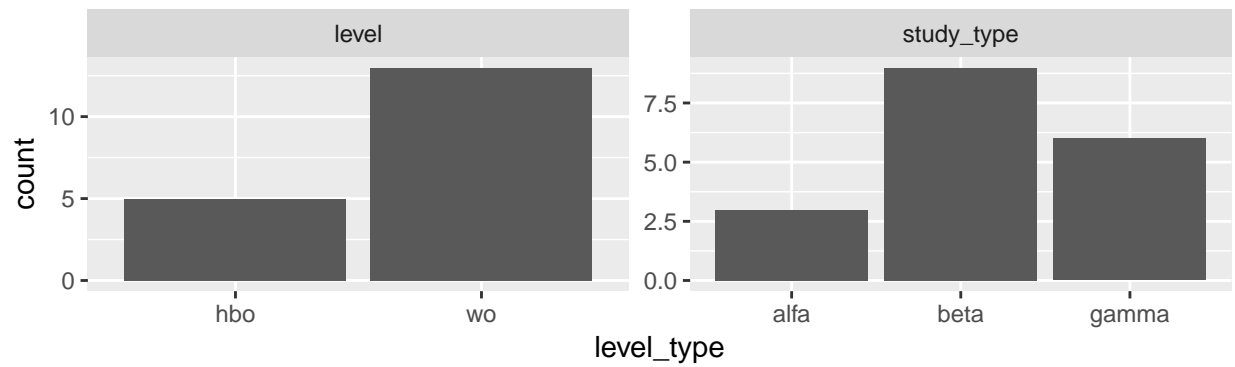


## Relatedness distribution

### Per `study_type`

The density distribution is similar for the different groups. This seems to confirm what has been noted earlier: the different groups do not seem to give an explanation for different types of relatedness rating sidtributions.
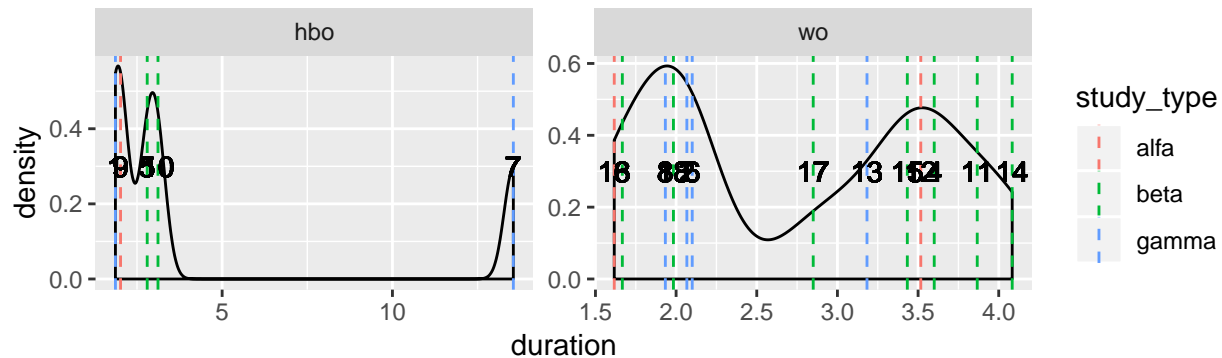


### Per `level`

The WO participants have given slightly more extreme rating than HBO participants.
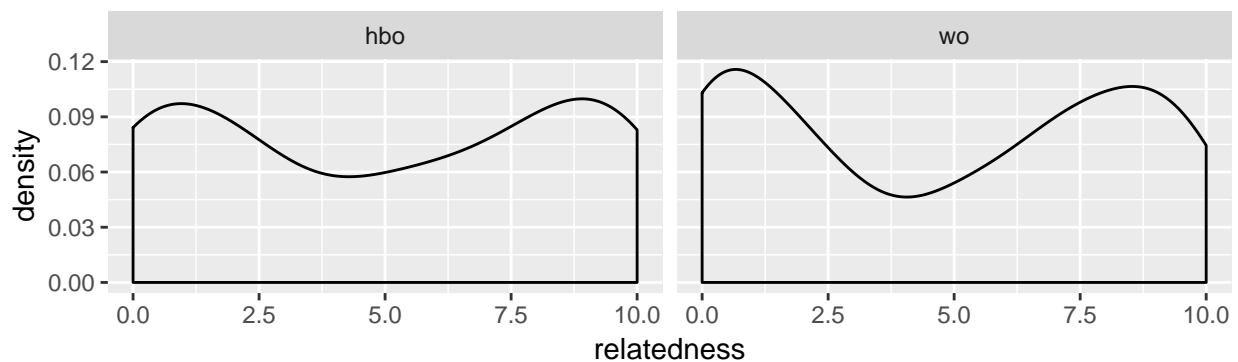


## Duration denisty

Participant 7 took very long to complete the questionnaire in comparison with the others.

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



## Correlations

### Between participants

```
## Joining, by = "id"
```

In the table below we can see that high correlations, above .9, occur between participants of different groups, and not only between participants of the same group.

| id | participant2 | correlation | study_type1 | level1 | study_type2 | level2 |
|----|----|----|----|----|----|----|
| 1 | 4 | 0.9150917 | gamma | hbo | beta | wo |
| 10 | 9 | 0.9017205 | beta | hbo | alfa | hbo |
| 11 | 9 | 0.9159699 | beta | wo | alfa | hbo |
| 16 | 4 | 0.9266538 | alfa | wo | beta | wo |

Low correlations, of lower than .5 occur only between one specific one specific participant, participant 7, and two others. Participant 7 is one of the two participants with a gamma and hbo education. Additionally, we have seen earlier that participant 7 took relatively long to complete the form.

| id | participant2 | correlation | study_type1 | level1 | study_type2 | level2 |
|----|----|----|----|----|----|----|
| 1 | 7 | 0.4814133 | gamma | hbo | gamma | hbo |
| 12 | 7 | 0.4960998 | alfa | wo | gamma | hbo |

| id | level | study_type | profession | duration |
|----|----|----|----|----|
| 1 | hbo | gamma | HBO Digital Communication and Media | 1.866667 mins |
| 7 | hbo | gamma | Psychology | 13.550000 mins |

## Does a model that includes `study_type` explain the variance better than one that does not?

We will use an `lmer` model from the `lme4` package, because we we are interested in the effects of `study_type`, and not the individual differences that we can not control for. The intercept of each individual (`id`) will be defined as a random effect. We will leave out the `NA`'s.

Furthermore, `word_pair` will be defined as a random effect. At this moment we are mainly interested in whether `study_type` explains differences in relatedness in general. We don't want to look at the effect of

`study_type` on the relatedness ratings of each `word_pair`.

### The null-model

The null-model only includes the intercept for each `word_pair` and the intercept of each individual as a random effects.

```
df_naomit <- df %>% na.omit

model0 <- lmer(relatedness ~ (1|word_pair) + (1|id), data = df_naomit, REML = F)
```

### The alternative model (includes `study_type`)

In the table we can see that there is a relatively large difference between the ratings of people with a `gamma` background and those with an `alpha background`. `estimate > (Intercept)` gives us the relatedness intercept of relatedness of `people` with an `alpha` background. The `estimate` value at `study_typebeta` gives the difference with this intercept. The same holds for `study_typegamma`.

```
model1 <- lmer(relatedness ~ study_type + (1|word_pair) + (1|id), data = df_naomit, REML = F)

model1 %>%
  tidy() %>%
  rename("t-value" = statistic) %>%
  select(-group) %>%
  kable()
```

```
## Warning in bind_rows_(x, .id): binding factor and character vector,
## coercing into character vector
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

| term | estimate | std.error | t-value |
|------|---------:|----------:|--------:|
| (Intercept) | 4.6111111 | 0.6856217 | 6.7254448 |
| study_typebeta | 0.3777778 | 0.4538399 | 0.8324032 |
| study_typegamma | 0.2000000 | 0.4813699 | 0.4154809 |
| sd__(Intercept).word_pair | 3.0770075 | NA | NA |
| sd__(Intercept).id | 0.5958940 | NA | NA |
| sd__Observation.Residual | 1.8028672 | NA | NA |

### Does the alternate model explain significantly more of the variance?

The p-value indicates that a model that includes `study_type` does not explain significantly more of the variance.

```
anova(model0, model1)
```

```
## Data: df_naomit
## Models:
## model0: relatedness ~ (1 | word_pair) + (1 | id)
## model1: relatedness ~ study_type + (1 | word_pair) + (1 | id)
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model0  4 2321.8 2339.0 -1156.9   2313.8
## model1  6 2325.1 2350.8 -1156.5   2313.1 0.7368      2     0.6918
```

## How about `level` and `duration` and combinations of those and `study_type`?

```
model2 <- lmer(relatedness ~ level + (1|word_pair) + (1|id), data = df_naomit, REML = F)
model3 <- lmer(relatedness ~ duration + (1|word_pair) + (1|id), data = df_naomit, REML = F)
```

level

```
## Data: df_naomit
## Models:
## model0: relatedness ~ (1 | word_pair) + (1 | id)
## model2: relatedness ~ level + (1 | word_pair) + (1 | id)
##        Df    AIC  BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model0  4 2321.8 2339 -1156.9   2313.8
## model2  5 2323.5 2345 -1156.8   2313.5 0.3398      1     0.5599
```

duration

```
## Data: df_naomit
## Models:
## model0: relatedness ~ (1 | word_pair) + (1 | id)
## model3: relatedness ~ duration + (1 | word_pair) + (1 | id)
##        Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## model0  4 2321.8 2339.0 -1156.9   2313.8
## model3  5 2323.4 2344.9 -1156.7   2313.4 0.406      1      0.524
```

Models including `level` xor `duration` do not explain significantly more of the variance than the null-model.