IXN 1 — Data Validation

ANONYMIZED

2023-09-08

Contents


```r
raw_data <- read_csv("data/wrangled_utterances_representations.csv")

#WRANGLE
df_insights <- raw_data %>%
  #rename and factorize columns
  mutate(
    sid = factor(UID), #NOT actually a unique utterance id, treat as sheet order id
    pid = factor(PID),
    DATASET = factor(recode(Notebook, "Happiness"="happiness", "Space"="space")), #cleanup diff case
    outcomeType = recode(DATASET, "happiness"="numeric", "space"="nominal"),
    top_code = factor(highlevel),
    mid_code = factor(`Data Type`),
    low_code = factor(UtteranceType),
    timestamp = Timestamp,
    repns = group,
    ixn = factor(interaction_used),
    utterance = Utterance,
    TASK = factor(recode(Condition, "Static"="static", "Interactive"="ixn" )),
    uid = factor(as.numeric(factor(paste(pid,factor(utterance))))) #construct a unique ID for utterance
  ) %>% select( #select only needed columns
    sid,uid, pid, TASK, DATASET, ixn, top_code, mid_code, low_code, repns, timestamp, utterance
```

```
  )

#DF OF UNIQUE UTTERANCES
df_uniques <- df_insights %>% select(uid, pid, TASK, DATASET) %>%
  distinct()  #take only unique utterances

glimpse(df_insights)
```

```
## Rows: 743
## Columns: 12
## $ sid       <fct> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ uid       <fct> 419, 431, 473, 449, 423, 446, 421, 477, 443, 457, 476, 451, ~
## $ pid       <fct> j2719eertu2, j2719eertu2, j2719eertu2, j2719eertu2, j2719eer~
## $ TASK      <fct> static, static, static, static, static, static, static, stat~
## $ DATASET   <fct> space, space, space, space, space, space, space, space, spac~
## $ ixn       <fct> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
## $ top_code  <fct> DATASET, DATASET, DATASET, VARIABLE, DATASET, VARIABLE, VARI~
## $ mid_code  <fct> NA, NA, NA, distribution (categorical), NA, distribution (ca~
## $ low_code  <fct> "data orientation", "data orientation", "data size", "distri~
## $ repns     <chr> "dataframe", "dataframe", "profile", "profile", "profile", "~
## $ timestamp <time> 19:57:00, 20:10:00, 21:27:00, 21:29:00, 21:38:00, 21:44:00,~
## $ utterance <chr> "\"Alright, so every row is the passenger, their home planet~
```

```
glimpse(df_uniques)
```

```
## Rows: 662
## Columns: 4
## $ uid     <fct> 419, 431, 473, 449, 423, 446, 421, 477, 443, 457, 476, 451, 48~
## $ pid     <fct> j2719eertu2, j2719eertu2, j2719eertu2, j2719eertu2, j2719eertu~
## $ TASK    <fct> static, static, static, static, static, static, static, static~
## $ DATASET <fct> space, space, space, space, space, space, space, space, space,~
```

DATA PROFILE

TODO TALK WITH DYLAN - resolve missing data in TASK, outcomeType, timestamp… are these the result of 'exploded' utterances that were dual coded? <– need to carry the other attributes across both obs - max of 2 detail-codes applied, correct? - where are the flag codes?

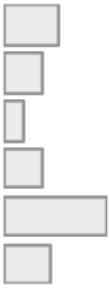
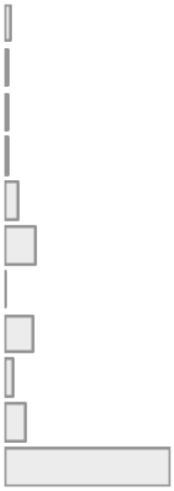
```
df_insights %>% summarytools::dfSummary(
             plain.ascii  = FALSE,
             graph.magnif = 0.75,
             style        = "grid",
             tmp.img.dir  = "temp",
             missing.col = FALSE,
             method = "render"
)
```

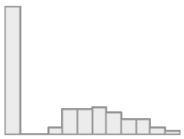
Data Frame Summary

df_insights Dimensions: 743 x 12
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	sid [factor]	1. 0 2. 1 3. 2 4. 3 5. 4 6. 5 7. 6 8. 7 9. 8 10. 9 [733 others]	1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 733 (98.7%)		743 (100.0%)	0 (0.0%)
2	uid [factor]	1. 1 2. 2 3. 3 4. 4 5. 5 6. 6 7. 7 8. 8 9. 9 10. 10 [652 others]	2 (0.3%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 2 (0.3%) 1 (0.1%) 1 (0.1%) 731 (98.4%)		743 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
3	pid [factor]	1. 3r2sh20ei 2. 4728sjuiz 3. 7382kwtue 4. 7ACC0B75 5. 8v892iige 6. 92ghd48xe 7. bjs827ee1u 8. E1D39056 9. iurmer289 10. j2719eertu2 [3 others]	103 (14.0%) 43 (5.9%) 54 (7.4%) 28 (3.8%) 49 (6.7%) 56 (7.6%) 29 (4.0%) 25 (3.4%) 87 (11.9%) 78 (10.6%) 182 (24.8%)		734 (98.8%)	9 (1.2%)
4	TASK [factor]	1. ixn 2. static	342 (46.6%) 392 (53.4%)		734 (98.8%)	9 (1.2%)
5	DATASET [factor]	1. happiness 2. space	420 (57.2%) 314 (42.8%)		734 (98.8%)	9 (1.2%)
6	ixn [factor]	1. FALSE 2. TRUE	634 (85.3%) 109 (14.7%)		743 (100.0%)	0 (0.0%)
7	top_code [factor]	1. ANALYSIS PROCESS 2. DATASET 3. RELATIONSHIP 4. VARIABLE	160 (21.5%) 176 (23.7%) 285 (38.4%) 122 (16.4%)		743 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
8	mid_code [factor]	1. distribution (continuous 2. distribution (categorical 3. relationship (categorical 4. relationship (categorical 5. relationship (continuous 6. relationship (multivariat	77 (18.0%) 54 (12.6%) 28 (6.6%) 55 (12.9%) 146 (34.2%) 67 (15.7%)		427 (57.5%)	316 (42.5%)
9	low_code [factor]	1. data orientation 2. data provenance 3. data size 4. distribution outlier (var 5. distribution range [min, 6. distribution shape [shape 7. distribution variance (sd 8. missing data 9. outlier (relationship) 10. plan of action [8 others]	16 (2.2%) 11 (1.5%) 9 (1.2%) 9 (1.2%) 33 (4.4%) 79 (10.6%) 1 (0.1%) 76 (10.2%) 20 (2.7%) 52 (7.0%) 437 (58.8%)		743 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
10	repns [character]	1. scatterplot 2. profile 3. none 4. dataframe 5. Multi-view Chart 6. data_dictionary 7. pairplot 8. lineplot 9. describe 10. double-profiler [14 others]	128 (17.2%) 107 (14.4%) 105 (14.1%) 74 (10.0%) 59 (7.9%) 56 (7.5%) 50 (6.7%) 36 (4.8%) 23 (3.1%) 23 (3.1%) 82 (11.0%)		743 (100.0%)	0 (0.0%)
11	timestamp [hms, difftime]	min : 868 med : 70710 max : 215160 units : secs	622 distinct values		738 (99.3%)	5 (0.7%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
12	utterance [character]	1. [Talking about the profil 2. actually, let me see if p 3. Although we have like les 4. And are they within range 5. And confidence in governm 6. And just I want to see ho 7. And so it looks like it s 8. And then if I had more ti 9. Because it does seem like 10. Data frame. Got a bunch o [652 others]	2 (0.3%) 2 (0.3%) 2 (0.3%) 2 (0.3%) 2 (0.3%) 2 (0.3%) 2 (0.3%) 2 (0.3%) 2 (0.3%) 2 (0.3%) 723 (97.3%)		743 (100.0%)	0 (0.0%)

DESCRIBE

```
#COUNTS
n_rows <- df_insights %>% nrow()
n_unique <- nlevels(df_insights$uid)
n_participants <- nlevels(df_insights$pid)

#count number of codes per unique utterance
s <- df_insights %>% group_by(uid) %>%
    dplyr::summarise(
      count = n()
) %>% arrange(desc(count), .by_group = TRUE)

max_codes <- max(s$count)

#display frequencies
(f <- freq(s$count,
     order    = "freq",
```

```
     rows     = 1:10,
     headings = FALSE))
```

```
## There are only 2 rows to show; higher numbers will be ignored
```

```
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1    581     87.76          87.76     87.76          87.76
##           2     81     12.24         100.00     12.24         100.00
##        <NA>      0                                0.00         100.00
##       Total    662    100.00         100.00    100.00         100.00
```

```
coded_single <- f[1,1]
coded_double <- f[2,1]
```

There are 743 coded utterances, representing 662 unique statements made by 13 in the study. 581 utterances were single-coded, while 81 utterances received two detail codes. No more than 2 were applied to any single utterance.

NUMBER OF UTTERANCES

How many utterances did each participant make in static (vs) interactive tasks?

```
#SUMMARY DFS
# utterances_by_participant <- df_uniques %>%
#   group_by(pid, TASK) %>%
#   dplyr::summarise(
#     n_utterances = n()
#   )


#SUMMARY TABLE
title = "Utterances by Participant and Task"
cols = c("Static Task","Interactive Task","Total Utterances")
cont <- table(df_uniques$pid, df_uniques$TASK)
cont %>% addmargins() %>% kbl(caption = title, col.names = cols) %>%  kable_classic()
```

```
# FOR DODGED NOT STACKED
# gf_bar(  pid ~ . , fill = ~ TASK, data = df_insights,
#            position = position_dodge(),
#            orientation = 'y') +

#UTTERANCES by PARTICPANT and TASK (horizontal)
gf_bar(  pid ~ uid , fill = ~ TASK, data = df_uniques) +
# %>%   gf_facet_grid(.~TASK) +
  labs(
    title = "Number of Utterances by Participant and Task",
    subtitle = "some participants were far more talkative than others",
    x = "number of unique utterances",
```

Table 2: Utterances by Participant and Task

	Static Task	Interactive Task	Total Utterances
3r2sh20ei	32	52	84
4728sjuiz	12	27	39
7382kwtue	18	22	40
7ACC0B75	9	17	26
8v892iige	21	24	45
92ghd48xe	28	27	55
bjs827ee1u	15	11	26
E1D39056	14	9	23
iurmer289	39	41	80
j2719eertu2	43	29	72
li832lin23	30	16	46
lkin27js09b	14	23	37
s294hoei	27	55	82
Sum	302	353	655

```
    y = "participant",
    fill = "Analysis Task"
  )
```

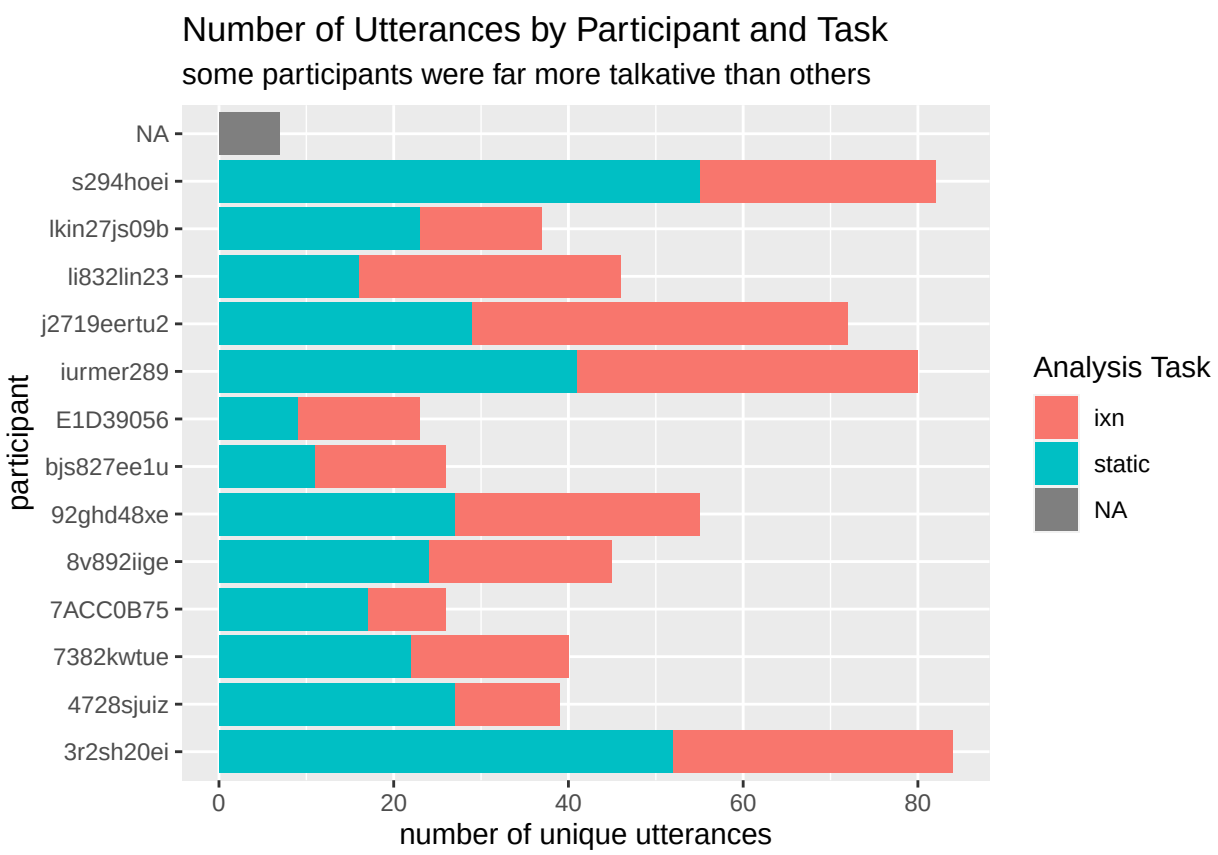
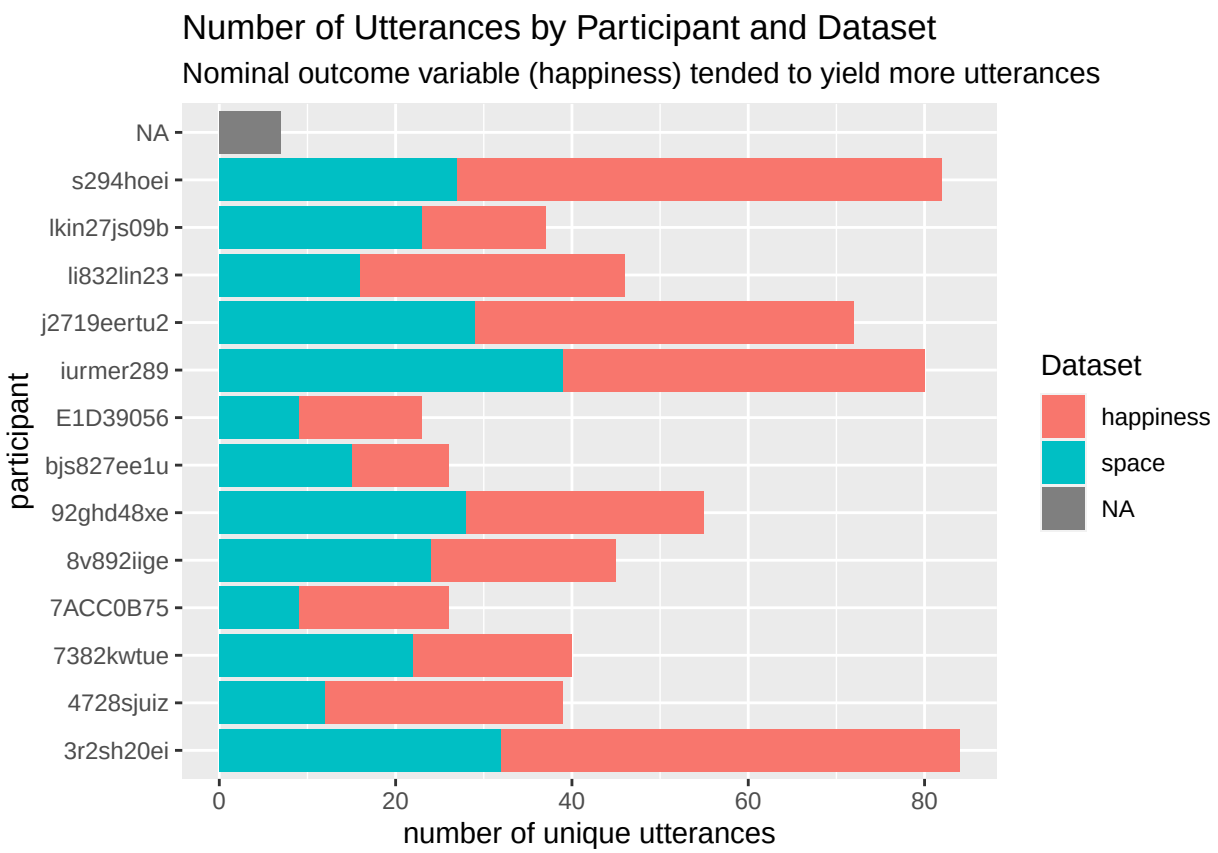


Table 3: Utterances by Participant and Dataset

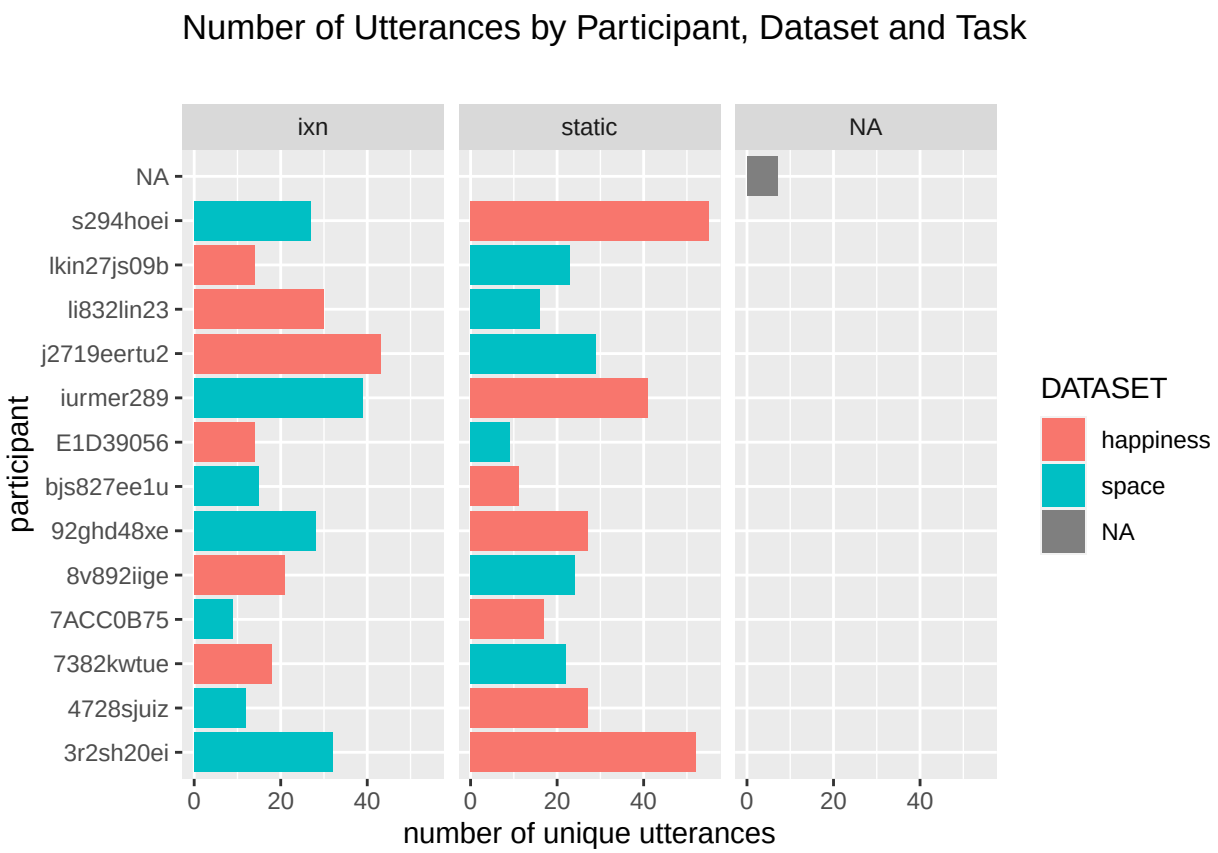
	Happiness (numeric)	Space (nominal)	Total Utterances
3r2sh20ei	52	32	84
4728sjuiz	27	12	39
7382kwtue	18	22	40
7ACC0B75	17	9	26
8v892iige	21	24	45
92ghd48xe	27	28	55
bjs827ee1u	11	15	26
E1D39056	14	9	23
iurmer289	41	39	80
j2719eertu2	43	29	72
li832lin23	30	16	46
lkin27js09b	14	23	37
s294hoei	55	27	82
Sum	370	285	655

```r
#SUMMARY TABLE
title = "Utterances by Participant and Dataset"
cols = c("Happiness (numeric)"," Space (nominal)","Total Utterances")
cont <- table(df_uniques$pid, df_uniques$DATASET)
cont %>% addmargins() %>% kbl(caption = title, col.names = cols) %>%  kable_classic()
```

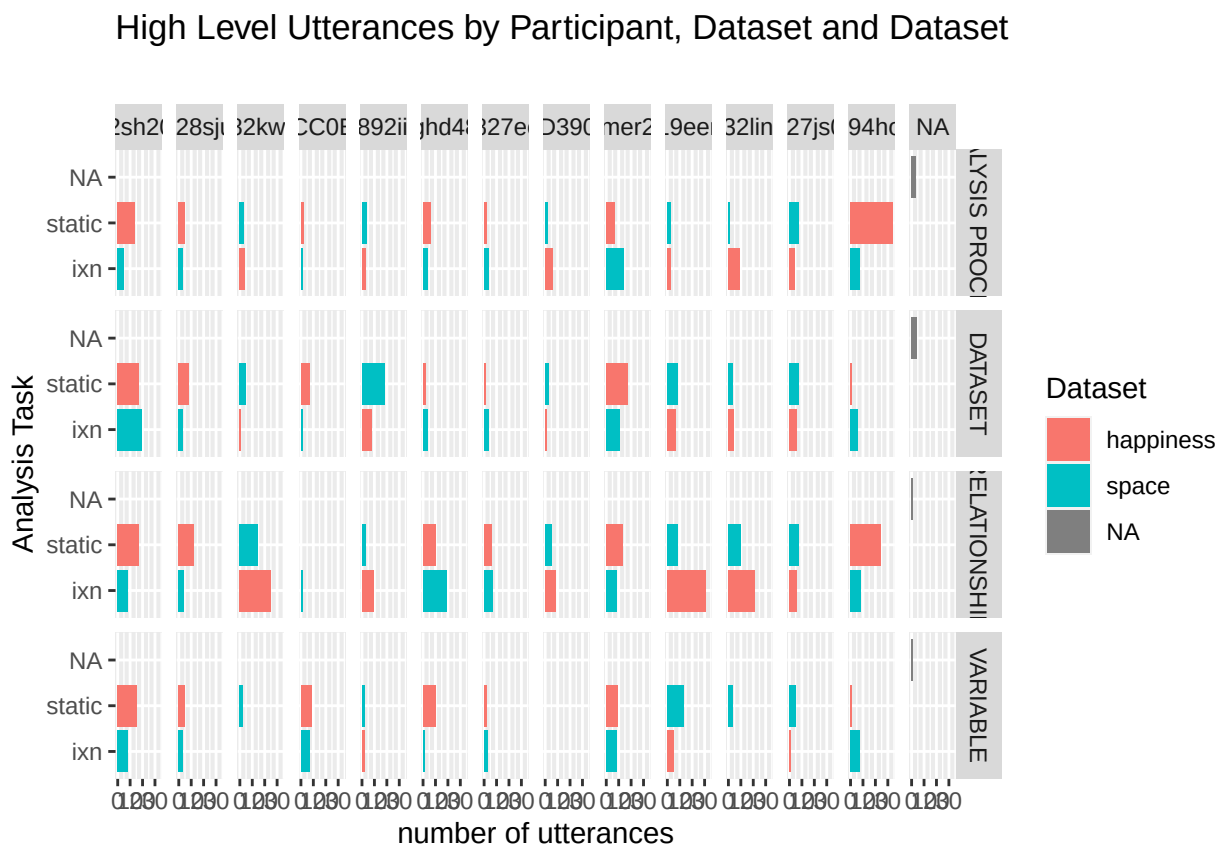
```r
# FOR DODGED NOT STACKED
# gf_bar(  pid ~ . , fill = ~ TASK, data = df_insights,
#            position = position_dodge(),
#            orientation = 'y') +

#UTTERANCES by PARTICPANT and DATASET (horizontal)
gf_bar(  pid ~ uid , fill = ~ DATASET, data = df_uniques) +
  # %>% gf_facet_grid(.~DATASET) +
  labs(
    title = "Number of Utterances by Participant and Dataset",
    subtitle = "Nominal outcome variable (happiness) tended to yield more utterances",
    x = "number of unique utterances",
    y = "participant",
    fill = "Dataset"
  )
```




```
#UTTERANCES by PARTICPANT and DATASET (horizontal)
gf_bar(  pid ~ uid , fill = ~ DATASET, data = df_uniques) %>%
  gf_facet_grid(.~TASK) +
  labs(
    title = "Number of Utterances by Participant, Dataset and Task",
    subtitle = "",
    x = "number of unique utterances",
    y = "participant",
    fill = "DATASET"
  )
```




```
#UTTERANCES by PARTICPANT, TASK, and DATASET (horizontal)
#FACETED BY PARTICIPANT AND TOP CODE
gf_bar(  TASK ~ uid , fill = ~ DATASET, data = df_insights) %>%
  gf_facet_grid(top_code ~ pid) +
  labs(
    title = "High Level Utterances by Participant, Dataset and Dataset",
    subtitle = "",
    x = "number of utterances",
    y = "Analysis Task",
    fill = "Dataset"
  )
```

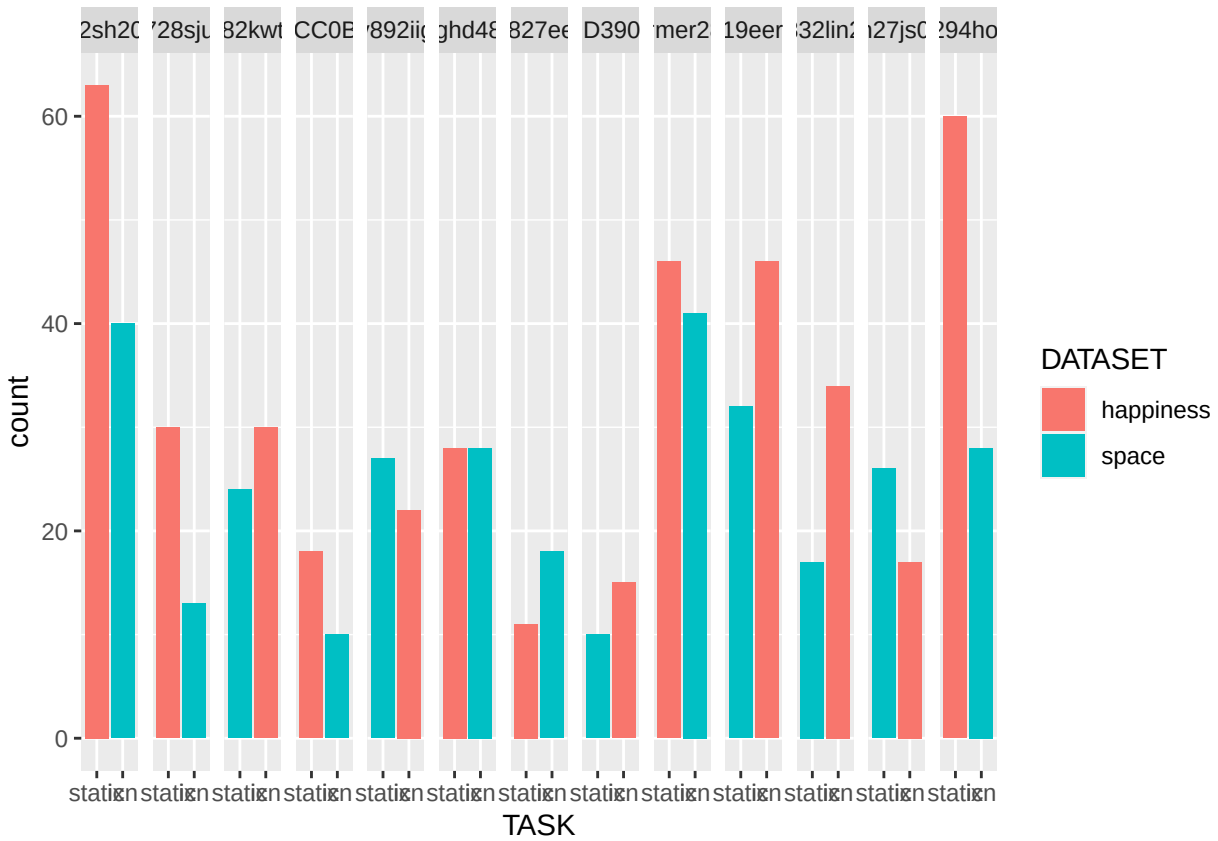


MODELLING

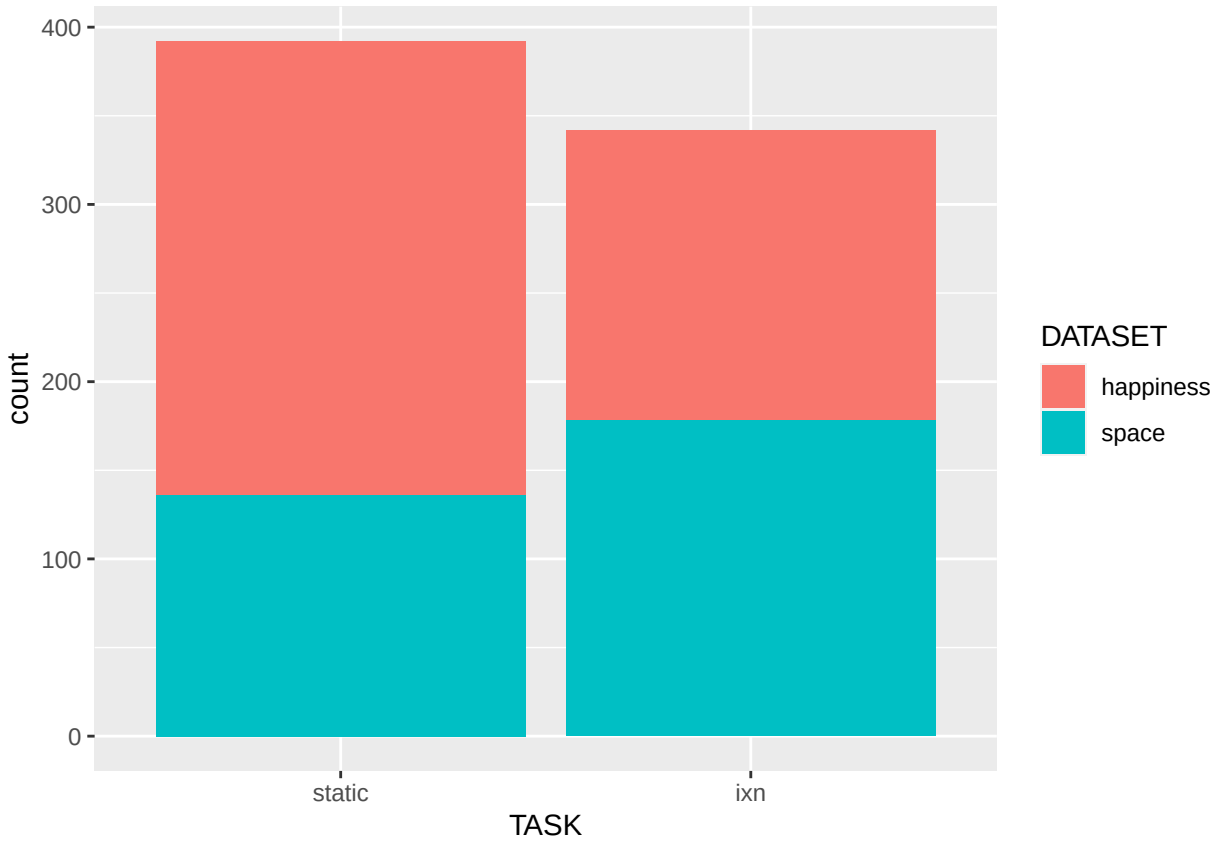
```
#DEFINE DATAFRAME
df_raw <- df_insights %>% select(pid, uid, TASK, DATASET) %>% mutate(
  TASK = factor(TASK, levels = c("static", "ixn")) #reorder factor levels
) %>% na.omit()
print("WARNING: THE FOLLOWING HAVE OMMITED MISSING DATA RATHER THAN FINDING THE SOURCE")
```

```
## [1] "WARNING: THE FOLLOWING HAVE OMMITED MISSING DATA RATHER THAN FINDING THE SOURCE"
```

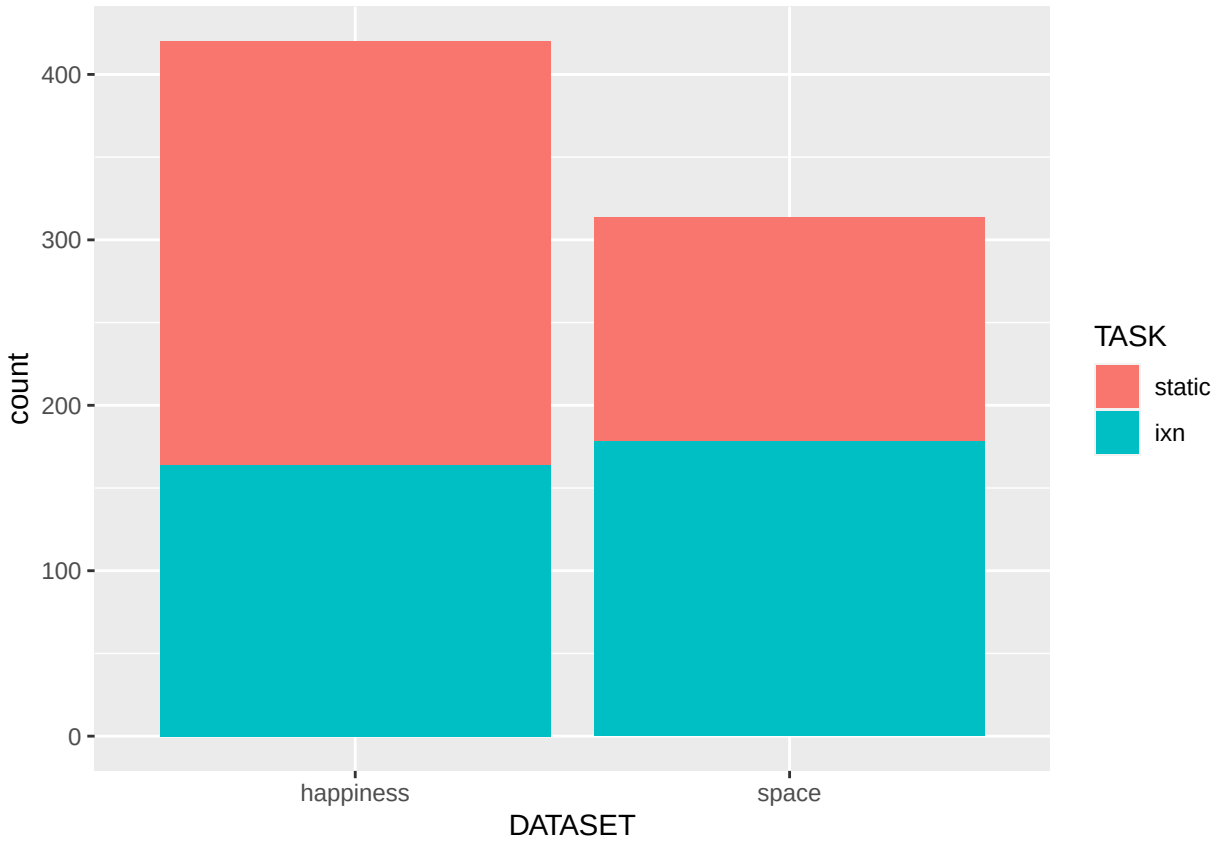
```
#DF SUMMARIZED BY SUBJECT
df_subject <- df_raw %>% group_by(pid, TASK, DATASET) %>% dplyr::summarise(
  n_utterances = n()
)

#VISUALIZE PARTICIPANTS
gf_bar( ~ TASK, fill = ~DATASET, data = df_raw) %>%
gf_facet_grid(.~pid)
```



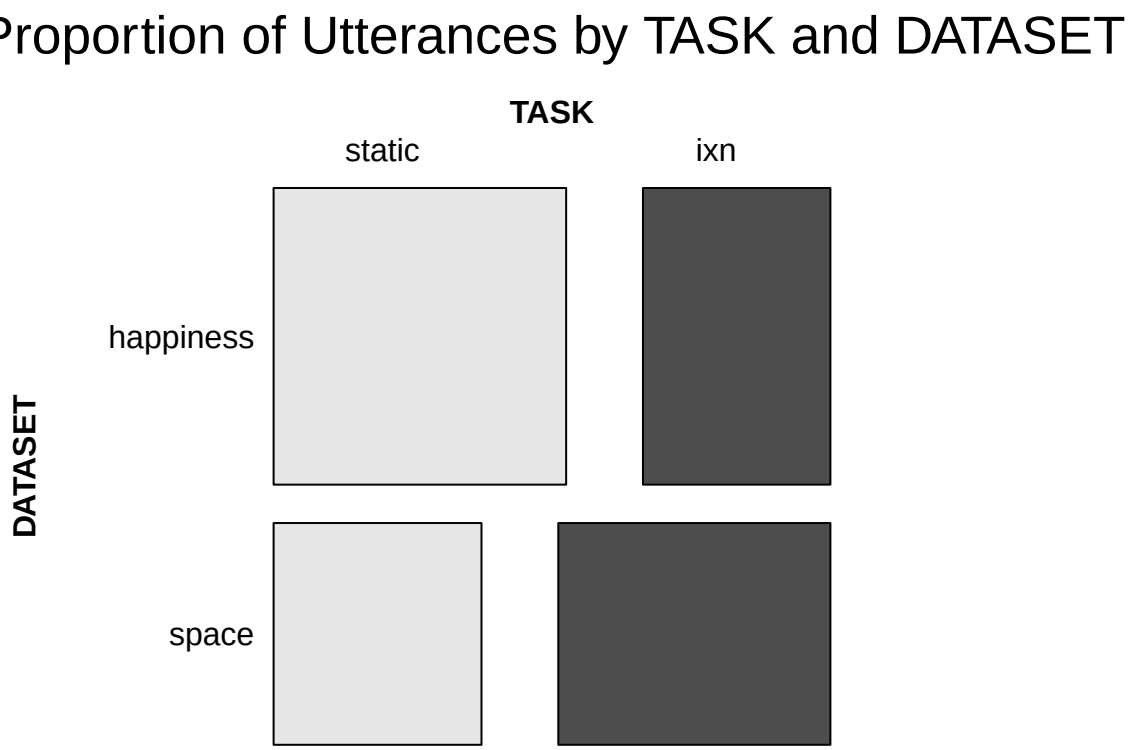

```
#VISUALIZE TOTALS
gf_bar (~ TASK, fill = ~DATASET, data = df_raw)
```



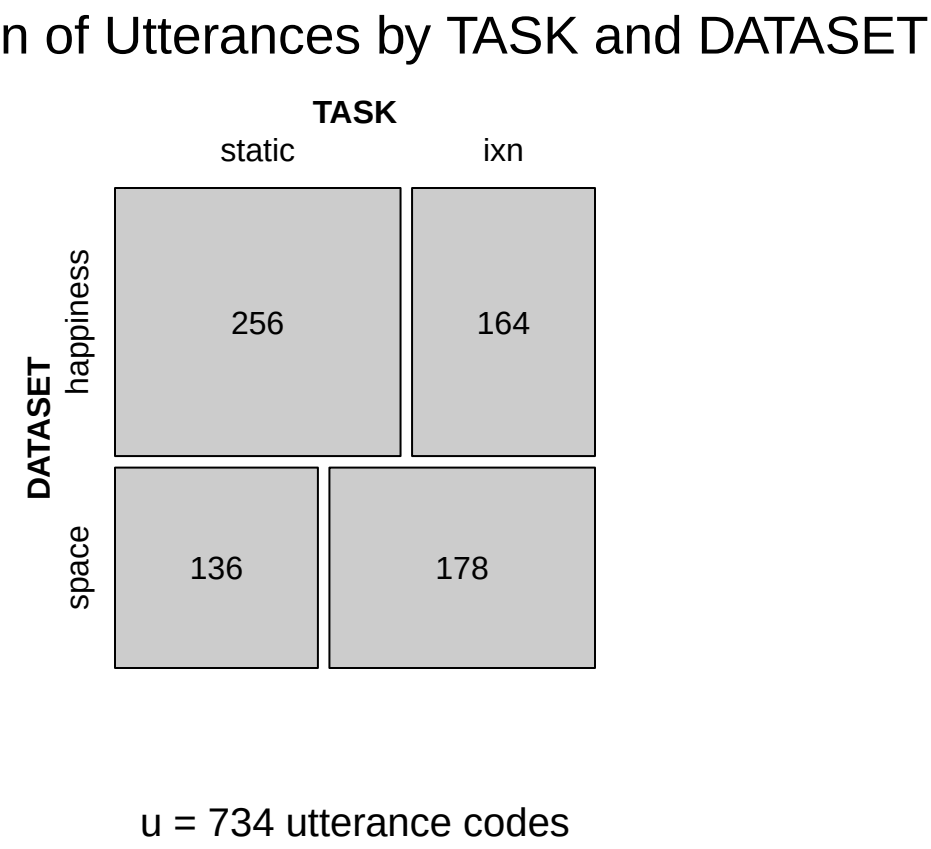

```
gf_bar (~ DATASET, fill = ~TASK, data = df_raw)
```




```
#MOSAIC PLOT
vcd::mosaic(main="Proportion of Utterances by TASK and DATASET",
            data = df_raw, TASK ~ DATASET, rot_labels=c(0,90,0,0),
            offset_varnames = c(left = 4.5), offset_labels = c(left = -0.5),just_labels = "right",
            spacing = spacing_dimequal(unit(1:2, "lines")))
```




```
mosaic(formula = ~DATASET + TASK,
       data = df_raw,
       main = "Proportion of Utterances by TASK and DATASET",
       sub = "u = 734 utterance codes",
       labeling = labeling_values,
       labeling_args = list(set_varnames = c(graph = "TASK",
                            datset = "DATASET")))
```



UTTERANCES by DATASET

How much variance in number of utterances is explained DATASET, TASK and PARTICIPANT?

OLS Fixed Effects Models

```
#NUMBER UTTERANCES predicted by DATASET + TASK --> OLS LINEAR REGRESSION
print("OLS-LM, UTTERANCES ~ DATASET + TASK")
```

```
## [1] "OLS-LM, UTTERANCES ~ DATASET + TASK"
```

```
m1 <- lm(n_utterances ~ DATASET + TASK, data = df_subject)
paste("Model")
```

```
## [1] "Model"
```

```
summ(m1)
```

```
paste("Partition Variance")
```

```
## [1] "Partition Variance"
```

Observations	26
Dependent variable	n_utterances
Type	OLS linear regression

F(2,23)	1.25
R^2	0.10
Adj. R^2	0.02

	Est.	S.E.	t val.	p
(Intercept)	33.80	4.69	7.21	0.00
DATASETspace	-7.90	5.56	-1.42	0.17
TASKixn	-3.24	5.56	-0.58	0.57

Standard errors: OLS

```
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: n_utterances
##           Df Sum Sq Mean Sq F value Pr(>F)
## DATASET    1  432.2  432.15  2.1614 0.1551
## TASK       1   67.8   67.75  0.3388 0.5662
## Residuals 23 4598.7  199.94
```

```
paste("Confidence Interval on Parameter Estimates")
```

```
## [1] "Confidence Interval on Parameter Estimates"
```

```
confint(m1)
```

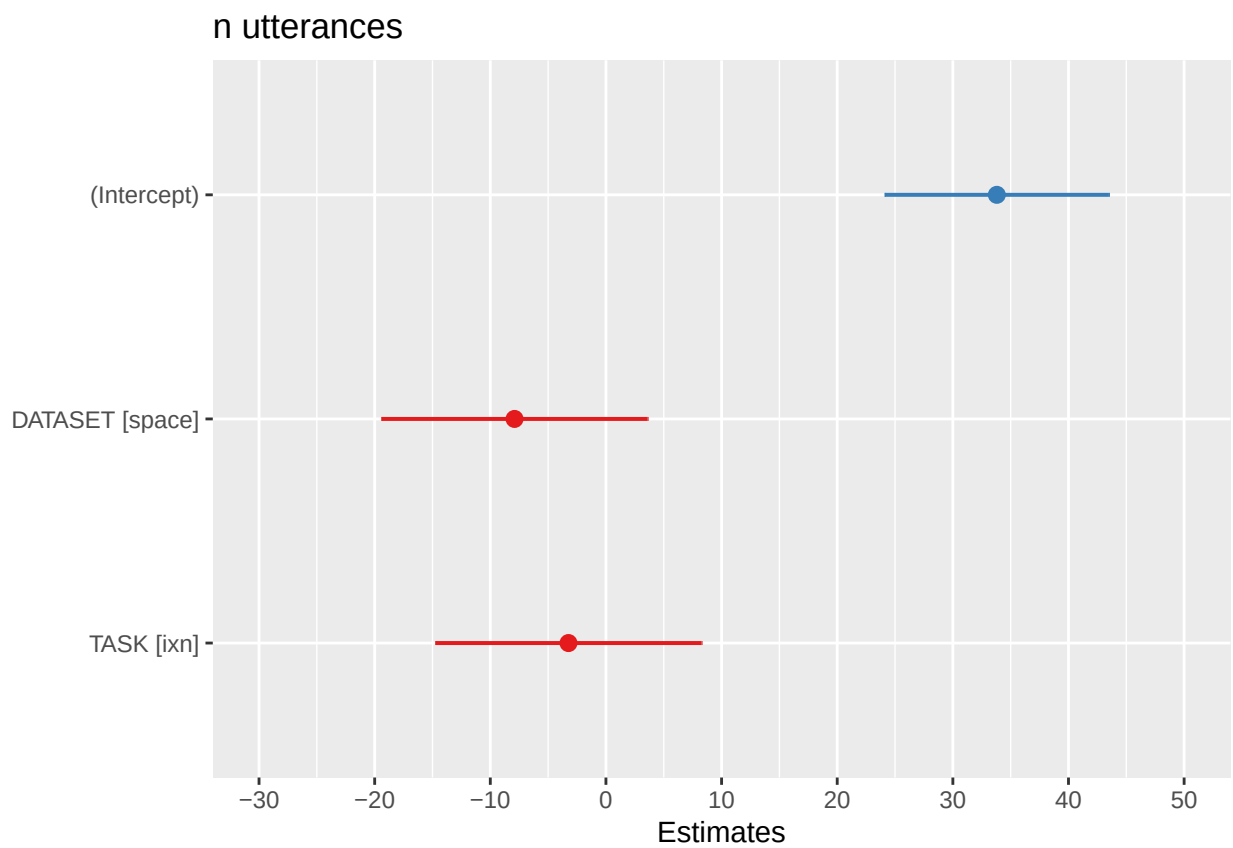
```
##                  2.5 %    97.5 %
## (Intercept)   24.10554 43.498855
## DATASETspace -19.41210  3.602573
## TASKixn      -14.74543  8.269240
```

```
report(m1) #sanity check
```

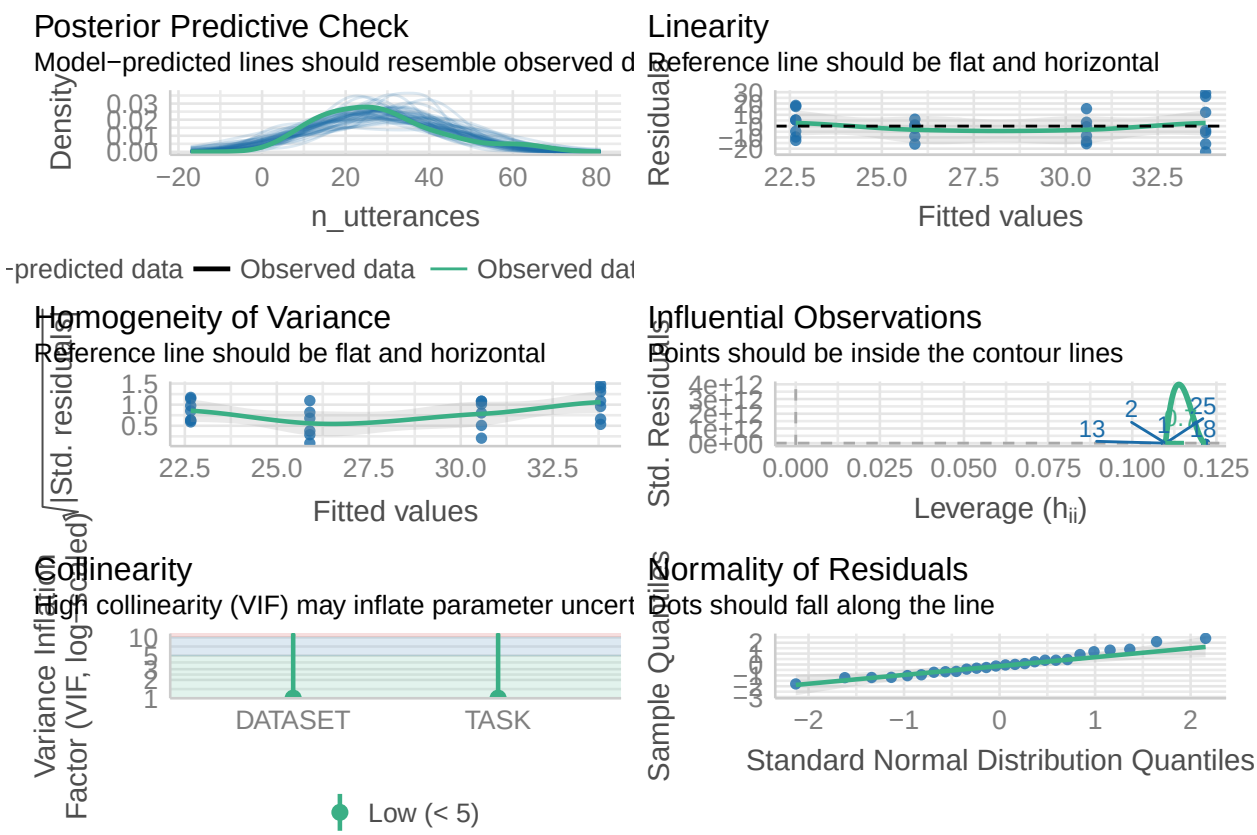
```
## We fitted a linear model (estimated using OLS) to predict n_utterances with
## DATASET and TASK (formula: n_utterances ~ DATASET + TASK). The model explains a
## statistically not significant and weak proportion of variance (R2 = 0.10, F(2,
## 23) = 1.25, p = 0.305, adj. R2 = 0.02). The model's intercept, corresponding to
## DATASET = happiness and TASK = static, is at 33.80 (95% CI [24.11, 43.50],
## t(23) = 7.21, p < .001). Within this model:
##
##   - The effect of DATASET [space] is statistically non-significant and negative
## (beta = -7.90, 95% CI [-19.41, 3.60], t(23) = -1.42, p = 0.169; Std. beta =
## -0.55, 95% CI [-1.36, 0.25])
##   - The effect of TASK [ixn] is statistically non-significant and negative (beta
```

```
## = -3.24, 95% CI [-14.75, 8.27], t(23) = -0.58, p = 0.566; Std. beta = -0.23,
## 95% CI [-1.03, 0.58])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

```
plot_model(m1,  show.intercept = TRUE)
```




```
check_model(m1)
```




```
#NUMBER UTTERANCES predicted by DATASET X TASK --> LINEAR REGRESSION
print("OLS-LM, UTTERANCES ~ DATASET * TASK")
```

```
## [1] "OLS-LM, UTTERANCES ~ DATASET * TASK"
```

```
m2 <- lm(n_utterances ~ DATASET * TASK, data = df_subject)
paste("Model")
```

```
## [1] "Model"
```

```
summ(m2)
```

Observations	26
Dependent variable	n_utterances
Type	OLS linear regression

F(3,22)	1.23
R^2	0.14
Adj. R^2	0.03

	Est.	S.E.	t val.	p
(Intercept)	36.57	5.32	6.87	0.00
DATASETspace	-13.90	7.84	-1.77	0.09
TASKixn	-9.24	7.84	-1.18	0.25
DATASETspace:TASKixn	12.00	11.08	1.08	0.29

Standard errors: OLS

```r
paste("Partition Variance")
```

```
## [1] "Partition Variance"
```

```r
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: n_utterances
##              Df Sum Sq Mean Sq F value Pr(>F)
## DATASET       1  432.2  432.15  2.1775 0.1542
## TASK          1   67.8   67.75  0.3414 0.5650
## DATASET:TASK  1  232.6  232.62  1.1721 0.2907
## Residuals    22 4366.1  198.46
```

```r
paste("Confidence Interval on Parameter Estimates")
```

```
## [1] "Confidence Interval on Parameter Estimates"
```

```r
confint(m2)
```

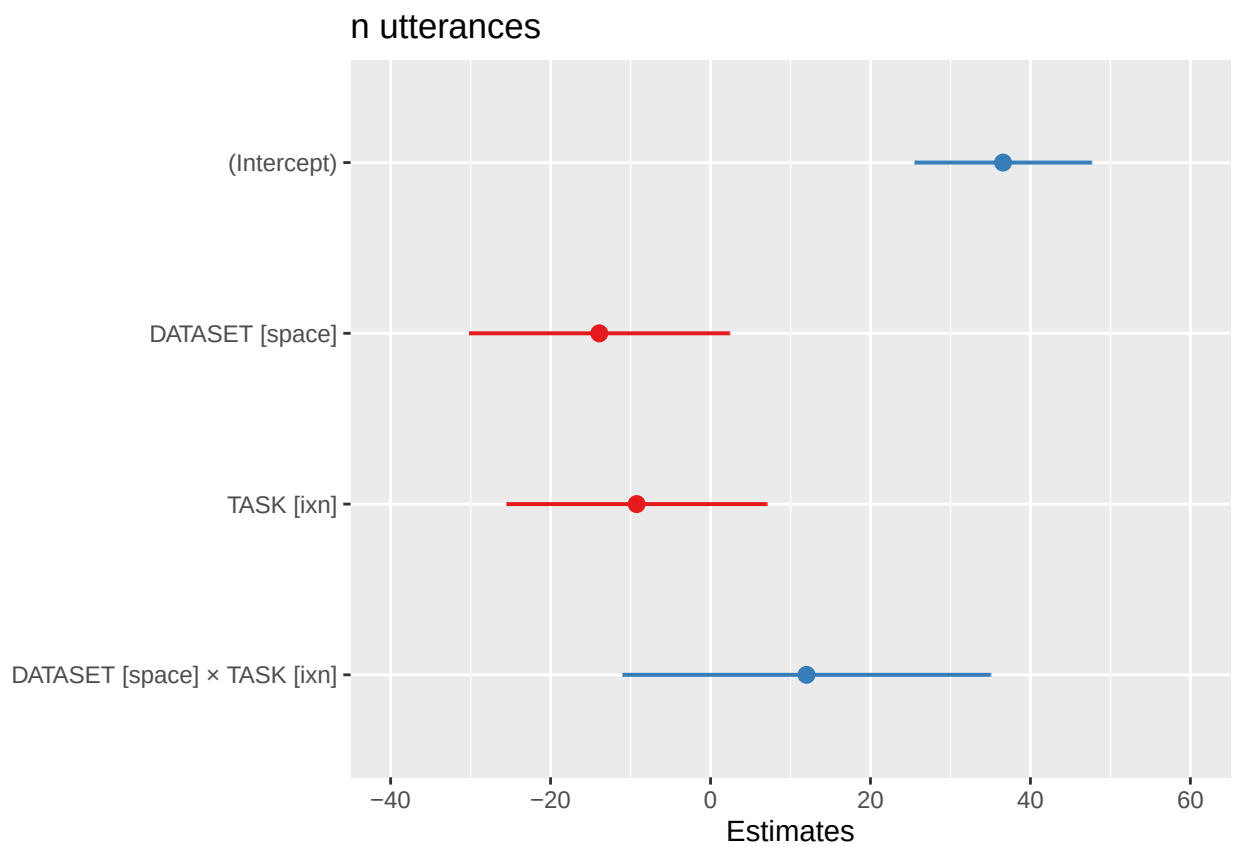
```
##                          2.5 %    97.5 %
## (Intercept)           25.52890 47.613954
## DATASETspace         -30.15892  2.349396
## TASKixn              -25.49225  7.016062
## DATASETspace:TASKixn -10.98685 34.986850
```

```r
report(m2) #sanity check
```

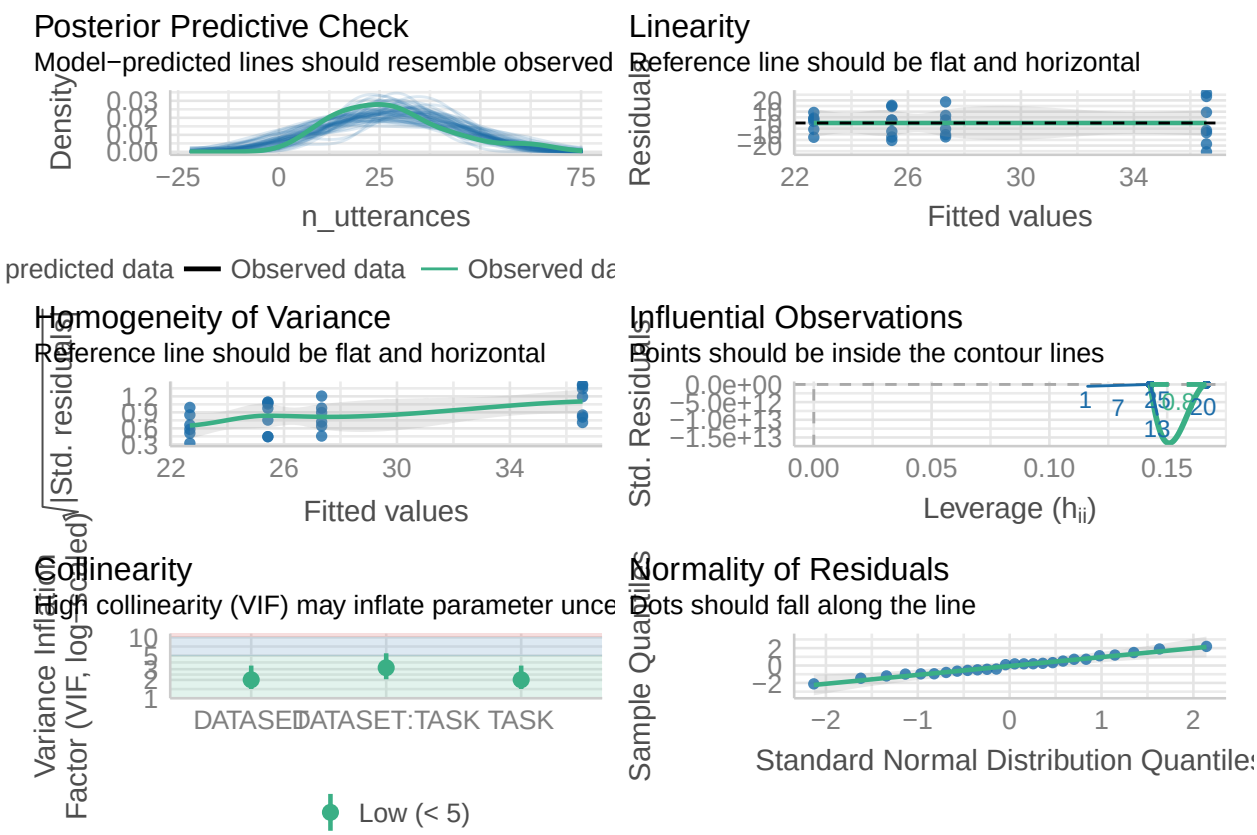
```
## We fitted a linear model (estimated using OLS) to predict n_utterances with
## DATASET and TASK (formula: n_utterances ~ DATASET * TASK). The model explains a
## statistically not significant and moderate proportion of variance (R2 = 0.14,
## F(3, 22) = 1.23, p = 0.322, adj. R2 = 0.03). The model's intercept,
## corresponding to DATASET = happiness and TASK = static, is at 36.57 (95% CI
## [25.53, 47.61], t(22) = 6.87, p < .001). Within this model:
##
##   - The effect of DATASET [space] is statistically non-significant and negative
## (beta = -13.90, 95% CI [-30.16, 2.35], t(22) = -1.77, p = 0.090; Std. beta =
## -0.97, 95% CI [-2.11, 0.16])
##   - The effect of TASK [ixn] is statistically non-significant and negative (beta
## = -9.24, 95% CI [-25.49, 7.02], t(22) = -1.18, p = 0.251; Std. beta = -0.65,
```

```
## 95% CI [-1.79, 0.49])
##   - The effect of DATASET [space] × TASK [ixn] is statistically non-significant
## and positive (beta = 12.00, 95% CI [-10.99, 34.99], t(22) = 1.08, p = 0.291;
## Std. beta = 0.84, 95% CI [-0.77, 2.45])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

```
plot_model(m2,  show.intercept = TRUE)
```




```
check_model(m2)
```



POISSON Fixed Effects Models

```
#NUMBER UTTERANCES predicted by DATASET + TASK --> POISSON DISTRIBUTION
print("GLM-POISSON, UTTERANCES ~ DATASET + TASK")
```

```
## [1] "GLM-POISSON, UTTERANCES ~ DATASET + TASK"
```

```
p.1 <- glm(n_utterances ~ DATASET + TASK, data = df_subject, family = "poisson")
paste("Model")
```

```
## [1] "Model"
```

```
summ(p.1)
```

Observations	26
Dependent variable	n_utterances
Type	Generalized linear model
Family	poisson
Link	log

$\chi^2(2)$	17.76
Pseudo-R² (Cragg-Uhler)	0.50
Pseudo-R² (McFadden)	0.06
AIC	292.63
BIC	296.41

	Est.	S.E.	z val.	p
(Intercept)	3.53	0.06	60.36	0.00
DATASETspace	-0.28	0.07	-3.77	0.00
TASKixn	-0.11	0.07	-1.55	0.12

Standard errors: MLE

```
paste("Partition Variance")
```

```
## [1] "Partition Variance"
```

```
anova(p.1)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: n_utterances
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev
## NULL                       25     172.78
## DATASET  1  15.3616        24     157.42
## TASK     1   2.4022        23     155.01
```

```
paste("Confidence Interval on Parameter Estimates")
```

```
## [1] "Confidence Interval on Parameter Estimates"
```

```
confint(p.1)
```

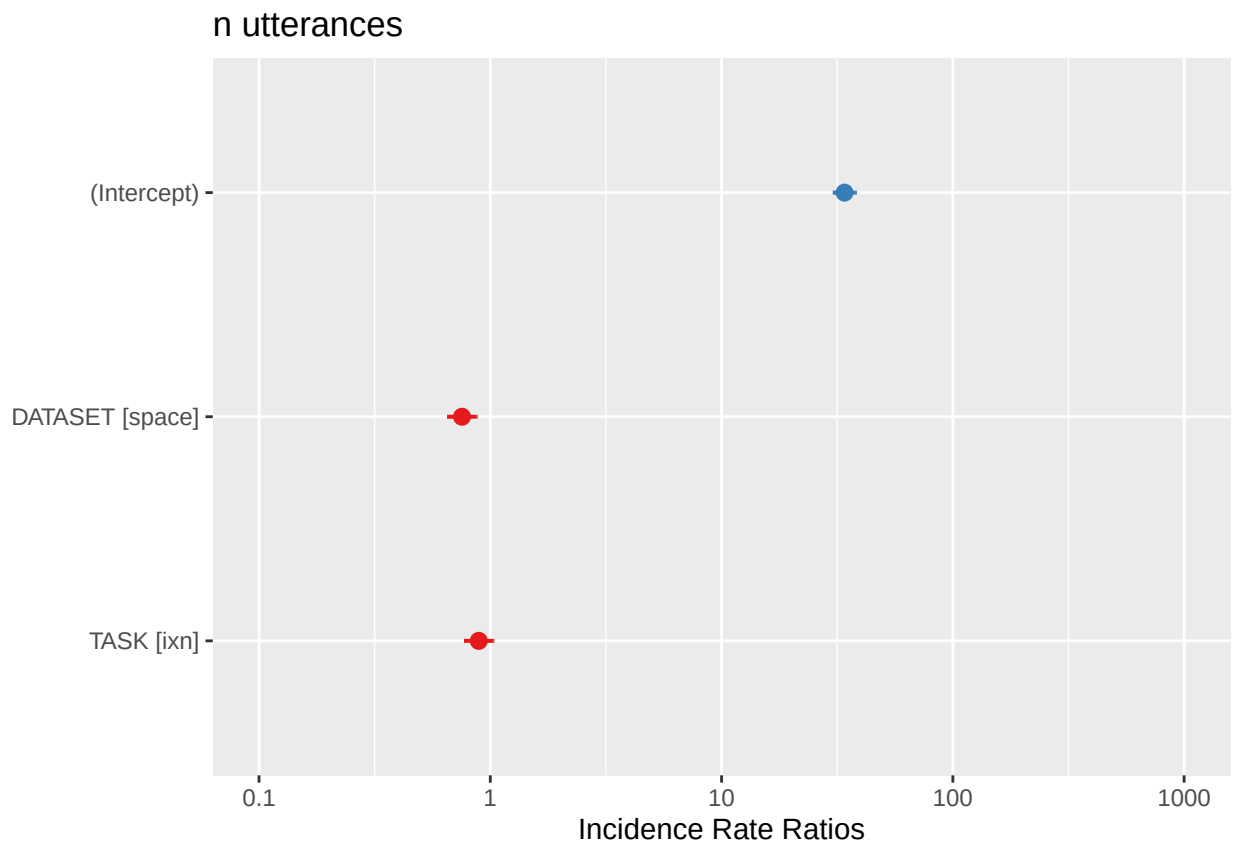
```
## Waiting for profiling to be done...
```

```
##                   2.5 %      97.5 %
## (Intercept)   3.4103351  3.63942450
## DATASETspace -0.4292558 -0.13583604
## TASKixn      -0.2606450  0.03037217
```

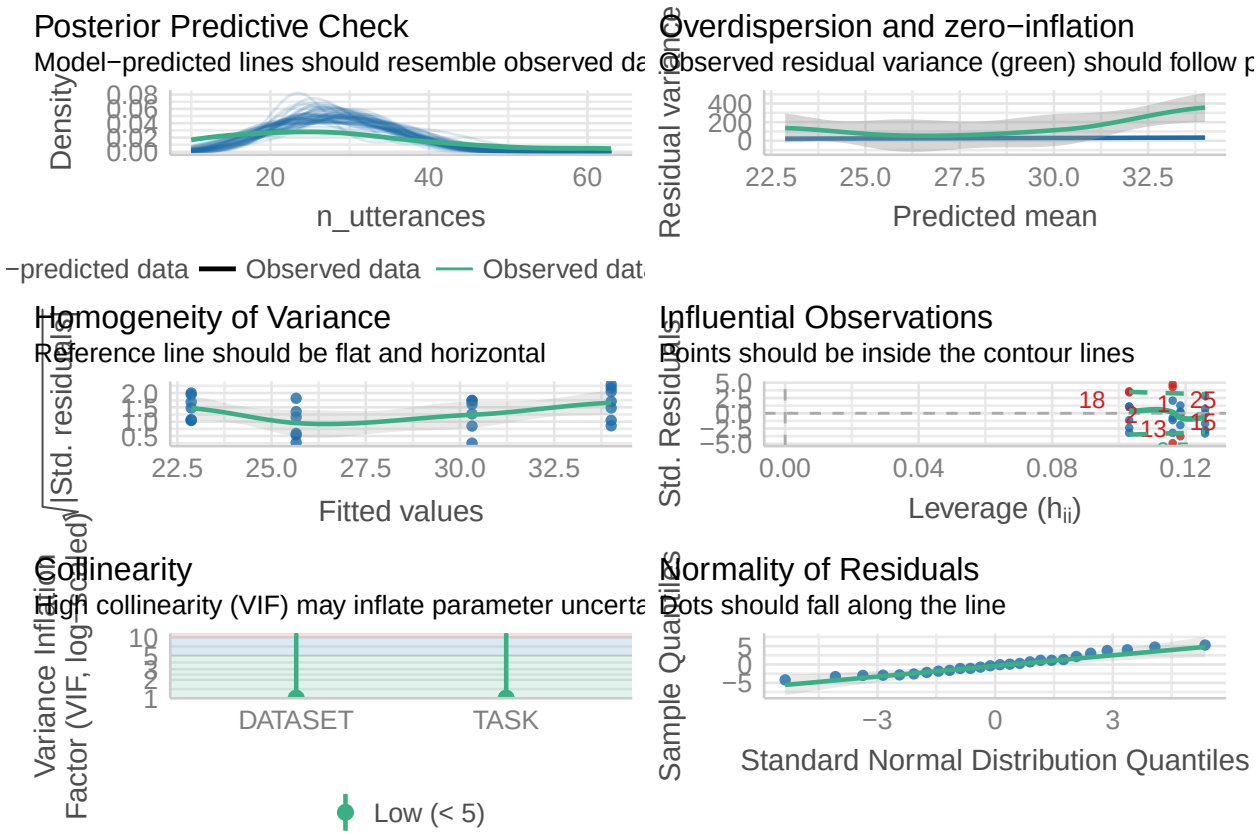
```
report(p.1) #sanity check
```

```
## We fitted a poisson model (estimated using ML) to predict n_utterances with
## DATASET and TASK (formula: n_utterances ~ DATASET + TASK). The model's
## explanatory power is substantial (Nagelkerke's R2 = 0.50). The model's
## intercept, corresponding to DATASET = happiness and TASK = static, is at 3.53
## (95% CI [3.41, 3.64], p < .001). Within this model:
##
##   - The effect of DATASET [space] is statistically significant and negative (beta
## = -0.28, 95% CI [-0.43, -0.14], p < .001; Std. beta = -0.28, 95% CI [-0.43,
## -0.14])
##   - The effect of TASK [ixn] is statistically non-significant and negative (beta
## = -0.11, 95% CI [-0.26, 0.03], p = 0.122; Std. beta = -0.11, 95% CI [-0.26,
## 0.03])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald z-distribution approximation.
```

```
plot_model(p.1,show.intercept = TRUE)
```




```
check_model(p.1)
```




```
#NUMBER UTTERANCES predicted by DATASET * TASK --> POISSON DISTRIBUTION
print("GLM-POISSON, UTTERANCES ~ DATASET X TASK")
```

```
## [1] "GLM-POISSON, UTTERANCES ~ DATASET X TASK"
```

```
p.2 <- glm(n_utterances ~ DATASET * TASK, data = df_subject, family = "poisson")
paste("Model")
```

```
## [1] "Model"
```

```
summ(p.2)
```

Observations	26
Dependent variable	n_utterances
Type	Generalized linear model
Family	poisson
Link	log

```
paste("Partition Variance")
```

```
## [1] "Partition Variance"
```

$\chi^2(3)$	25.01
Pseudo-R² (Cragg-Uhler)	0.62
Pseudo-R² (McFadden)	0.08
AIC	287.38
BIC	292.42

	Est.	S.E.	z val.	p
(Intercept)	3.60	0.06	57.59	0.00
DATASETspace	-0.48	0.11	-4.51	0.00
TASKixn	-0.29	0.10	-2.91	0.00
DATASETspace:TASKixn	0.41	0.15	2.68	0.01

Standard errors: MLE

```
anova(p.2)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: n_utterances
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                            25     172.78
## DATASET       1  15.3616        24     157.42
## TASK          1   2.4022        23     155.01
## DATASET:TASK  1   7.2491        22     147.77
```

```
paste("Confidence Interval on Parameter Estimates")
```

```
## [1] "Confidence Interval on Parameter Estimates"
```

```
confint(p.2)
```

```
## Waiting for profiling to be done...
```

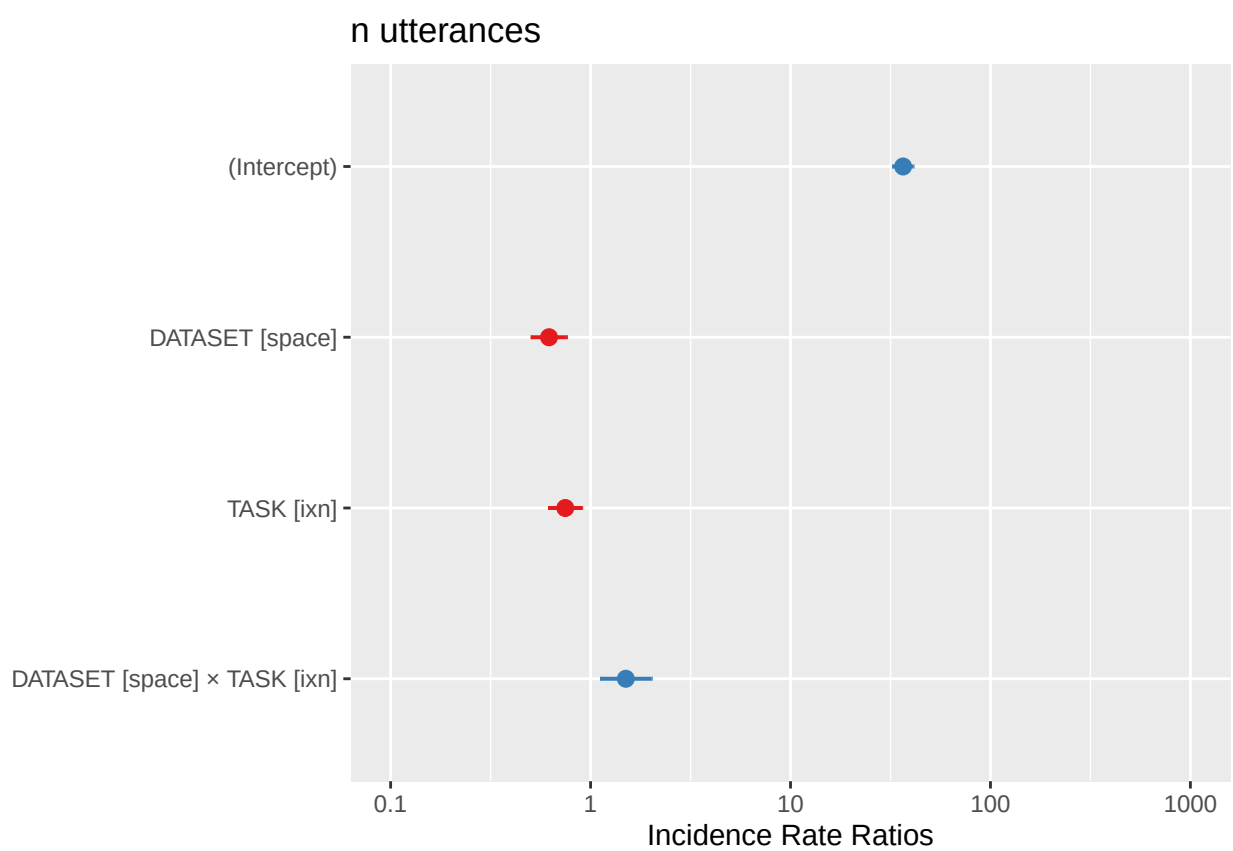
```
##                           2.5 %      97.5 %
## (Intercept)           3.4742131  3.71931074
## DATASETspace         -0.6887490 -0.27241989
## TASKixn              -0.4887608 -0.09637294
## DATASETspace:TASKixn  0.1101534  0.70460435
```

```
report(p.2) #sanity check
```

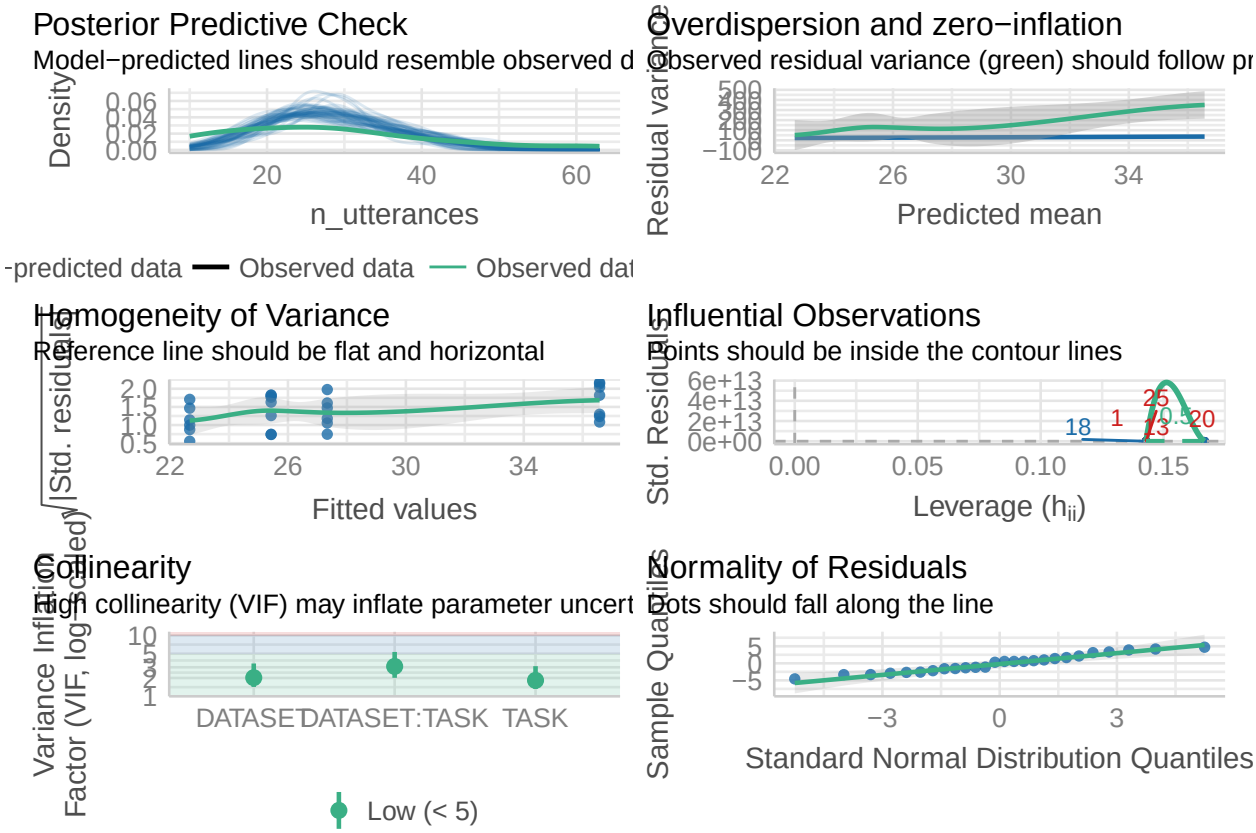
```
## We fitted a poisson model (estimated using ML) to predict n_utterances with
## DATASET and TASK (formula: n_utterances ~ DATASET * TASK). The model's
```

```
## explanatory power is substantial (Nagelkerke's R2 = 0.62). The model's
## intercept, corresponding to DATASET = happiness and TASK = static, is at 3.60
## (95% CI [3.47, 3.72], p < .001). Within this model:
##
##   - The effect of DATASET [space] is statistically significant and negative (beta
## = -0.48, 95% CI [-0.69, -0.27], p < .001; Std. beta = -0.48, 95% CI [-0.69,
## -0.27])
##   - The effect of TASK [ixn] is statistically significant and negative (beta =
## -0.29, 95% CI [-0.49, -0.10], p = 0.004; Std. beta = -0.29, 95% CI [-0.49,
## -0.10])
##   - The effect of DATASET [space] × TASK [ixn] is statistically significant and
## positive (beta = 0.41, 95% CI [0.11, 0.70], p = 0.007; Std. beta = 0.41, 95% CI
## [0.11, 0.70])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald z-distribution approximation.
```

```
plot_model(p.2,show.intercept = TRUE)
```



```r
check_model(p.2)
```



OLS Mixed Effects Models

```r
#NUMBER UTTERANCES predicted by DATASET + TASK | participatnt--> MIXED LINEAR REGRESSION
print("LMER, UTTERANCES ~ DATASET + TASK")
```

```
## [1] "LMER, UTTERANCES ~ DATASET + TASK"
```

```r
mm1 <- lmer(n_utterances ~ DATASET + TASK+ (1|pid), data = df_subject)
paste("Model")
```

```
## [1] "Model"
```

```r
summ(mm1)
```

Observations	26
Dependent variable	n_utterances
Type	Mixed effects linear regression

AIC	198.68
BIC	204.97
Pseudo-R^2 (fixed effects)	0.09
Pseudo-R^2 (total)	0.65

Fixed Effects					
	Est.	S.E.	t val.	d.f.	p
(Intercept)	33.80	4.17	8.10	20.51	0.00
DATASETspace	-7.90	3.39	-2.33	11.00	0.04
TASKixn	-3.24	3.39	-0.96	11.00	0.36

p values calculated using Satterthwaite d.f.

Random Effects		
Group	Parameter	Std. Dev.
pid	(Intercept)	10.98
Residual		8.61

Grouping Variables		
Group	# groups	ICC
pid	13	0.62

```
paste("Partition Variance")
```

```
## [1] "Partition Variance"
```

```
anova(mm1)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##         Sum Sq Mean Sq NumDF DenDF F value  Pr(>F)
## DATASET 403.75  403.75     1    11  5.4421 0.03967 *
## TASK     67.75   67.75     1    11  0.9132 0.35980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
paste("Confidence Interval on Parameter Estimates")
```

```
## [1] "Confidence Interval on Parameter Estimates"
```

```
confint(mm1)
```

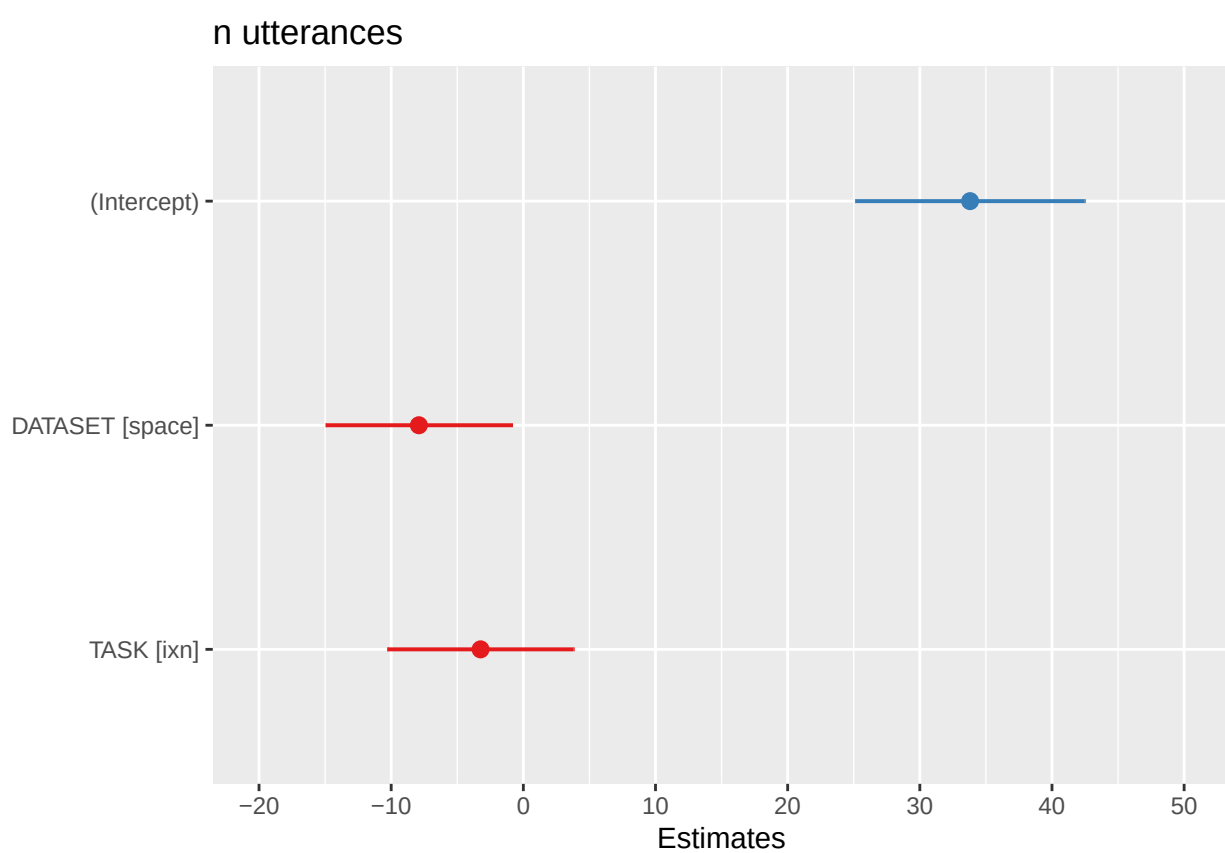
```
## Computing profile confidence intervals ...
```

```
##                   2.5 %    97.5 %
## .sig01         5.590803 17.904130
## .sigma         5.634678 12.309924
## (Intercept)   25.679890 41.924506
## DATASETspace -14.494243 -1.315280
## TASKixn       -9.827577  3.351386
```

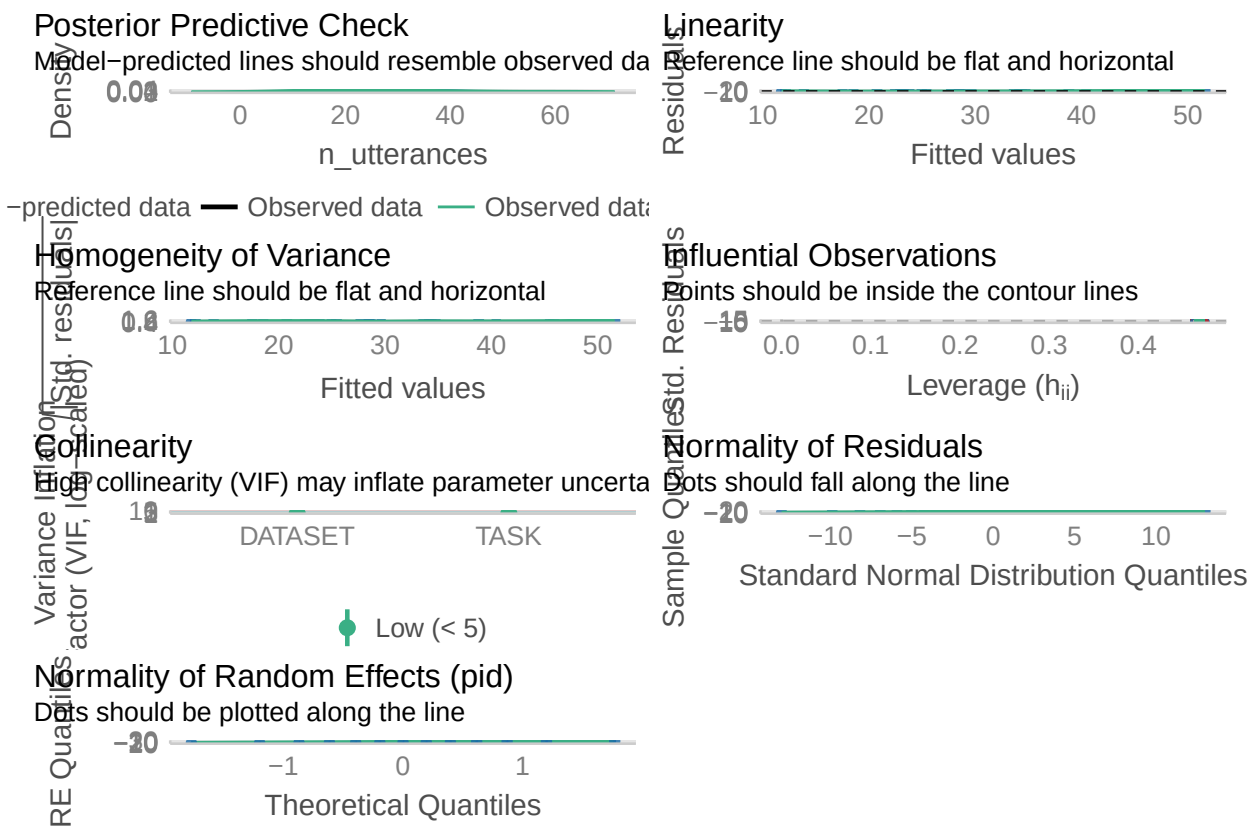
```r
report(mm1) #sanity check
```

```
## We fitted a linear mixed model (estimated using REML and nloptwrap optimizer)
## to predict n_utterances with DATASET and TASK (formula: n_utterances ~ DATASET
## + TASK). The model included pid as random effect (formula: ~1 | pid). The
## model's total explanatory power is substantial (conditional R2 = 0.65) and the
## part related to the fixed effects alone (marginal R2) is of 0.09. The model's
## intercept, corresponding to DATASET = happiness and TASK = static, is at 33.80
## (95% CI [25.12, 42.48], t(21) = 8.10, p < .001). Within this model:
##
##   - The effect of DATASET [space] is statistically significant and negative (beta
## = -7.90, 95% CI [-14.95, -0.86], t(21) = -2.33, p = 0.030; Std. beta = -0.55,
## 95% CI [-1.05, -0.06])
##   - The effect of TASK [ixn] is statistically non-significant and negative (beta
## = -3.24, 95% CI [-10.28, 3.81], t(21) = -0.96, p = 0.350; Std. beta = -0.23,
## 95% CI [-0.72, 0.27])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

```r
plot_model(mm1,  show.intercept = TRUE)
```



```r
check_model(mm1)
```




```r
#NUMBER UTTERANCES predicted by DATASET * TASK  | participatnt--> MIXED LINEAR REGRESSION
print("LMER, UTTERANCES ~ DATASET X TASK")
```

```
## [1] "LMER, UTTERANCES ~ DATASET X TASK"
```

```r
mm2 <- lmer(n_utterances ~ DATASET * TASK + (1|pid), data = df_subject)
paste("Model")
```

```
## [1] "Model"
```

```r
summ(mm2)
```

Observations	26
Dependent variable	n_utterances
Type	Mixed effects linear regression

```r
paste("Partition Variance")
```

```
## [1] "Partition Variance"
```

AIC	192.83
BIC	200.38
Pseudo-R² (fixed effects)	0.13
Pseudo-R² (total)	0.67

Fixed Effects					
	Est.	S.E.	t val.	d.f.	p
(Intercept)	36.57	5.32	6.87	15.80	0.00
DATASETspace	-13.90	7.84	-1.77	15.80	0.10
TASKixn	-9.24	7.84	-1.18	15.80	0.26
DATASETspace:TASKixn	12.00	14.13	0.85	11.00	0.41

p values calculated using Satterthwaite d.f.

Random Effects		
Group	Parameter	Std. Dev.
pid	(Intercept)	11.15
Residual		8.61

Grouping Variables		
Group	# groups	ICC
pid	13	0.63

```
anova(mm2)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##              Sum Sq Mean Sq NumDF DenDF F value  Pr(>F)
## DATASET      403.75  403.75     1    11  5.4421 0.03967 *
## TASK          67.75   67.75     1    11  0.9132 0.35980
## DATASET:TASK  53.48   53.48     1    11  0.7208 0.41398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
paste("Confidence Interval on Parameter Estimates")
```

```
## [1] "Confidence Interval on Parameter Estimates"
```

```
confint(mm2)
```

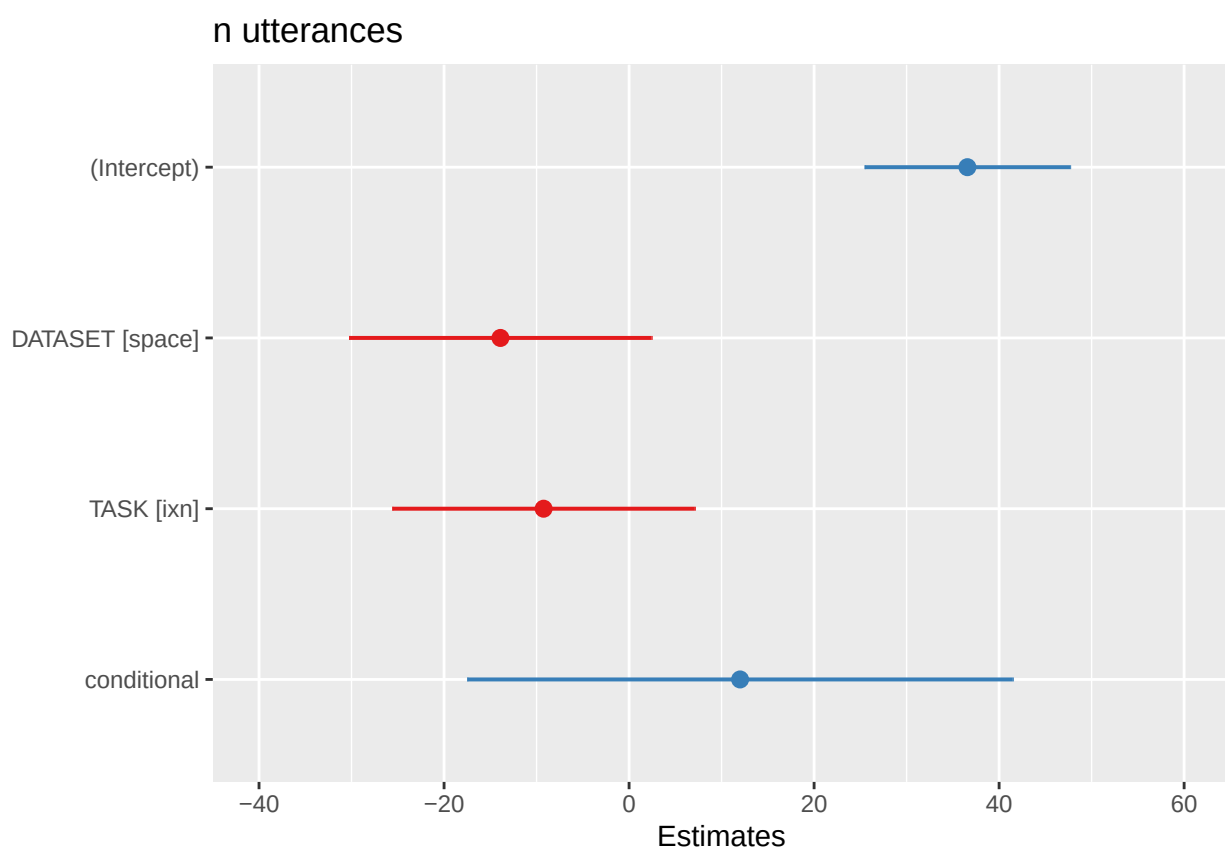
```
## Computing profile confidence intervals ...
```

```
##                            2.5 %    97.5 %
## .sig01                  5.088362 17.291859
## .sigma                  5.634705 12.310118
## (Intercept)            26.436694 46.706164
## DATASETspace          -28.822687  1.013164
## TASKixn               -24.156020  5.679831
## DATASETspace:TASKixn -15.486866 39.486867
```

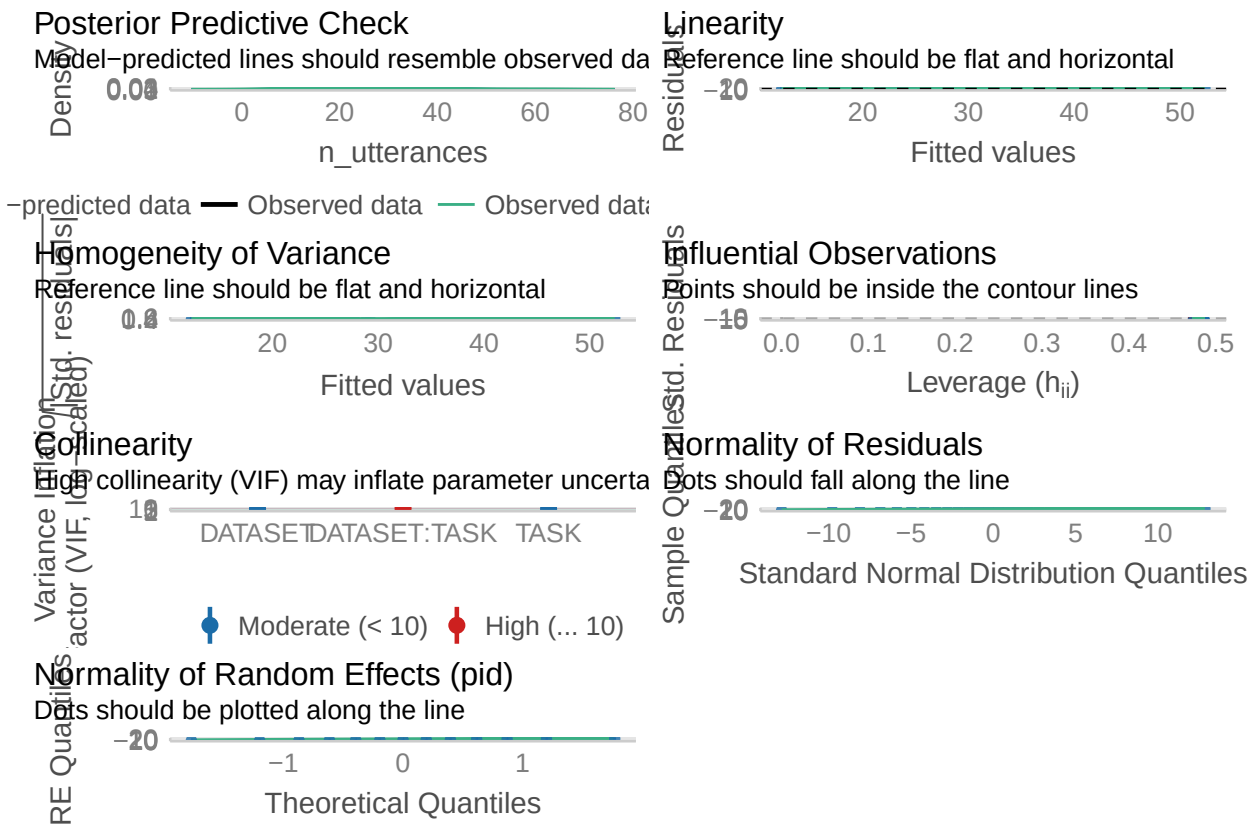
```r
report(mm2) #sanity check
```

```
## We fitted a linear mixed model (estimated using REML and nloptwrap optimizer)
## to predict n_utterances with DATASET and TASK (formula: n_utterances ~ DATASET
## * TASK). The model included pid as random effect (formula: ~1 | pid). The
## model's total explanatory power is substantial (conditional R2 = 0.67) and the
## part related to the fixed effects alone (marginal R2) is of 0.13. The model's
## intercept, corresponding to DATASET = happiness and TASK = static, is at 36.57
## (95% CI [25.46, 47.68], t(20) = 6.87, p < .001). Within this model:
##
##   - The effect of DATASET [space] is statistically non-significant and negative
## (beta = -13.90, 95% CI [-30.25, 2.44], t(20) = -1.77, p = 0.091; Std. beta =
## -0.97, 95% CI [-2.12, 0.17])
##   - The effect of TASK [ixn] is statistically non-significant and negative (beta
## = -9.24, 95% CI [-25.59, 7.11], t(20) = -1.18, p = 0.252; Std. beta = -0.65,
## 95% CI [-1.79, 0.50])
##   - The effect of DATASET [space] × TASK [ixn] is statistically non-significant
## and positive (beta = 12.00, 95% CI [-17.48, 41.48], t(20) = 0.85, p = 0.406;
## Std. beta = 0.84, 95% CI [-1.22, 2.90])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

```r
plot_model(mm2,  show.intercept = TRUE)
```



```
check_model(mm2)
```



POISSON Mixed Effects Models

```
#NUMBER UTTERANCES predicted by TASK + DATASET  | participatnt--> POISSON MIXED LINEAR REGRESSION
print("POISSON-MER, UTTERANCES ~ DATASET + TASK")
```

```
## [1] "POISSON-MER, UTTERANCES ~ DATASET + TASK"
```

```
pmm1 <- glmer(n_utterances ~ TASK + DATASET + (1|pid), data = df_subject, family = "poisson")
paste("Model")
```

```
## [1] "Model"
```

```
summ(pmm1)
```

```
paste("Partition Variance")
```

```
## [1] "Partition Variance"
```

Observations	26
Dependent variable	n_utterances
Type	Mixed effects generalized linear model
Family	poisson
Link	log

AIC	206.46
BIC	211.49
Pseudo-R^2 (fixed effects)	0.10
Pseudo-R^2 (total)	0.84

Fixed Effects				
	Est.	S.E.	z val.	p
(Intercept)	3.43	0.13	26.57	0.00
TASKixn	-0.09	0.08	-1.19	0.23
DATASETspace	-0.28	0.08	-3.65	0.00

Random Effects		
Group	Parameter	Std. Dev.
pid	(Intercept)	0.41

Grouping Variables		
Group	# groups	ICC
pid	13	0.15

```
anova(pmm1)
```

```
## Analysis of Variance Table
##         npar  Sum Sq Mean Sq F value
## TASK       1  3.2811  3.2811  3.2811
## DATASET    1 13.2458 13.2458 13.2458
```

```
paste("Confidence Interval on Parameter Estimates")
```

```
## [1] "Confidence Interval on Parameter Estimates"
```

```
confint(pmm1)
```

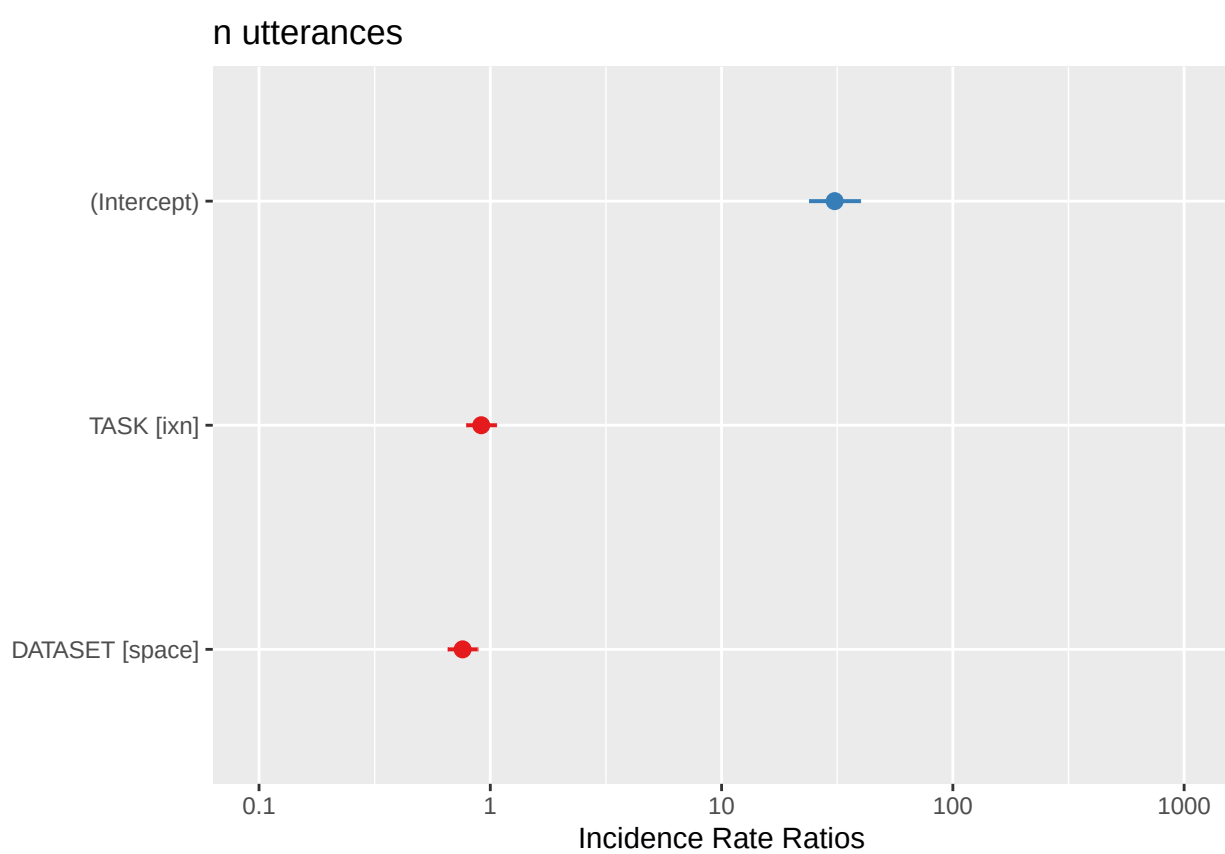
```
## Computing profile confidence intervals ...
```

```
##                   2.5 %      97.5 %
## .sig01        0.2746462  0.66576686
## (Intercept)   3.1540729  3.69253649
## TASKixn      -0.2383290  0.05831842
## DATASETspace -0.4245031 -0.12764249
```

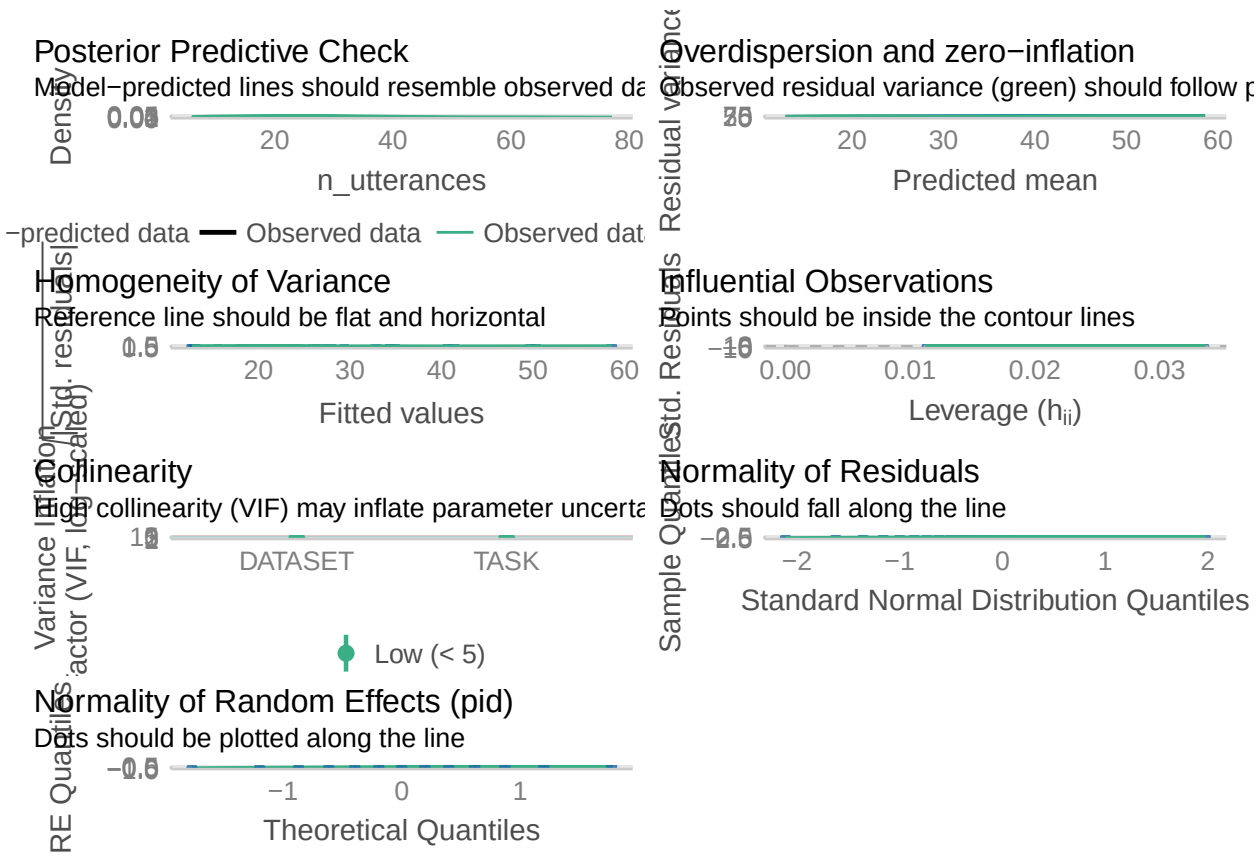
```r
report(pmm1) #sanity check
```

```
## We fitted a poisson mixed model (estimated using ML and Nelder-Mead optimizer)
## to predict n_utterances with TASK and DATASET (formula: n_utterances ~ TASK +
## DATASET). The model included pid as random effect (formula: ~1 | pid). The
## model's total explanatory power is substantial (conditional R2 = 0.84) and the
## part related to the fixed effects alone (marginal R2) is of 0.10. The model's
## intercept, corresponding to TASK = static and DATASET = happiness, is at 3.43
## (95% CI [3.18, 3.68], p < .001). Within this model:
##
##   - The effect of TASK [ixn] is statistically non-significant and negative (beta
## = -0.09, 95% CI [-0.24, 0.06], p = 0.232; Std. beta = -0.09, 95% CI [-0.24,
## 0.06])
##   - The effect of DATASET [space] is statistically significant and negative (beta
## = -0.28, 95% CI [-0.42, -0.13], p < .001; Std. beta = -0.28, 95% CI [-0.42,
## -0.13])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald z-distribution approximation.
```

```r
plot_model(pmm1,  show.intercept = TRUE)
```



```r
check_model(pmm1)
```




```r
#NUMBER UTTERANCES predicted by TASK X DATASET  | participatnt--> POISSON MIXED LINEAR REGRESSION
print("POISSON-MER, UTTERANCES ~ DATASET X TASK")
```

```
## [1] "POISSON-MER, UTTERANCES ~ DATASET X TASK"
```

```r
pmm2 <- glmer(n_utterances ~ TASK * DATASET + (1|pid), data = df_subject, family = "poisson")
paste("Model")
```

```
## [1] "Model"
```

```r
summ(pmm2)
```

Observations	26
Dependent variable	n_utterances
Type	Mixed effects generalized linear model
Family	poisson
Link	log

AIC	208.07
BIC	214.36
Pseudo-R² (fixed effects)	0.12
Pseudo-R² (total)	0.84

Fixed Effects				
	Est.	S.E.	z val.	p
(Intercept)	3.50	0.17	20.99	0.00
TASKixn	-0.24	0.25	-0.97	0.33
DATASETspace	-0.43	0.25	-1.71	0.09
TASKixn:DATASETspace	0.30	0.48	0.63	0.53

Random Effects		
Group	Parameter	Std. Dev.
pid	(Intercept)	0.40

Grouping Variables		
Group	# groups	ICC
pid	13	0.14

```r
paste("Partition Variance")
```

```
## [1] "Partition Variance"
```

```r
anova(pmm2)
```

```
## Analysis of Variance Table
##               npar  Sum Sq Mean Sq F value
## TASK             1  3.3827  3.3827  3.3827
## DATASET          1 13.2136 13.2136 13.2136
## TASK:DATASET     1  0.4015  0.4015  0.4015
```

```r
paste("Confidence Interval on Parameter Estimates")
```

```
## [1] "Confidence Interval on Parameter Estimates"
```

```r
confint(pmm2)
```

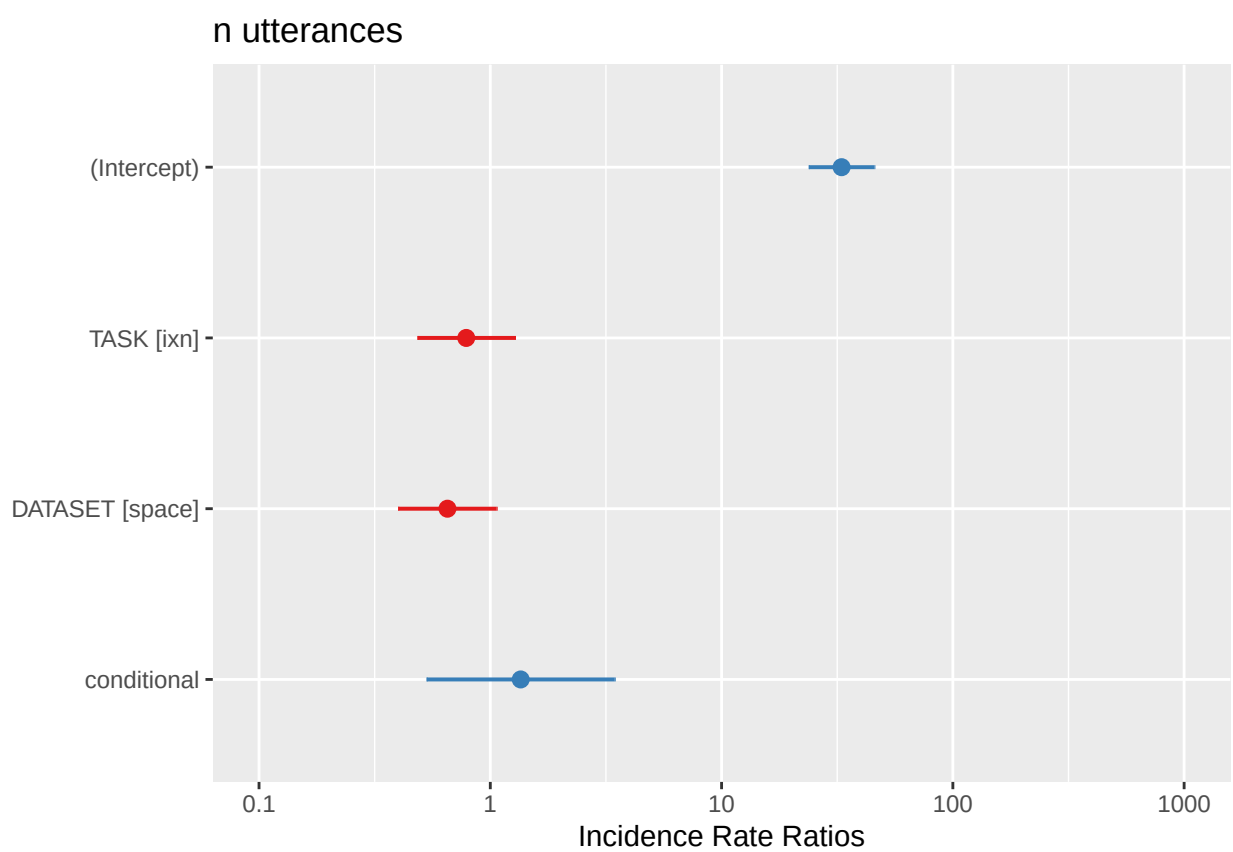
```
## Computing profile confidence intervals ...
```

```
##                           2.5 %    97.5 %
## .sig01                0.2676997 0.6545456
## (Intercept)           3.1401510 3.8418542
## TASKixn              -0.7570693 0.2834377
## DATASETspace         -0.9484807 0.1001263
## TASKixn:DATASETspace -0.7157996 1.3097794
```

```r
report(pmm2) #sanity check
```

```
## We fitted a poisson mixed model (estimated using ML and Nelder-Mead optimizer)
## to predict n_utterances with TASK and DATASET (formula: n_utterances ~ TASK *
## DATASET). The model included pid as random effect (formula: ~1 | pid). The
## model's total explanatory power is substantial (conditional R2 = 0.84) and the
## part related to the fixed effects alone (marginal R2) is of 0.12. The model's
## intercept, corresponding to TASK = static and DATASET = happiness, is at 3.50
## (95% CI [3.17, 3.82], p < .001). Within this model:
##
##   - The effect of TASK [ixn] is statistically non-significant and negative (beta
## = -0.24, 95% CI [-0.72, 0.25], p = 0.333; Std. beta = -0.24, 95% CI [-0.72,
## 0.25])
##   - The effect of DATASET [space] is statistically non-significant and negative
## (beta = -0.43, 95% CI [-0.92, 0.06], p = 0.088; Std. beta = -0.43, 95% CI
## [-0.92, 0.06])
##   - The effect of TASK [ixn] × DATASET [space] is statistically non-significant
## and positive (beta = 0.30, 95% CI [-0.63, 1.24], p = 0.526; Std. beta = 0.30,
## 95% CI [-0.63, 1.24])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald z-distribution approximation.
```

```r
plot_model(pmm2,  show.intercept = TRUE)
```



```
check_model(pmm2)
```