Comparative Analysis: ZSP vs FSP + RAG

Diego ALARCON, Donato GENTILE

1. Data Loading and Preparation

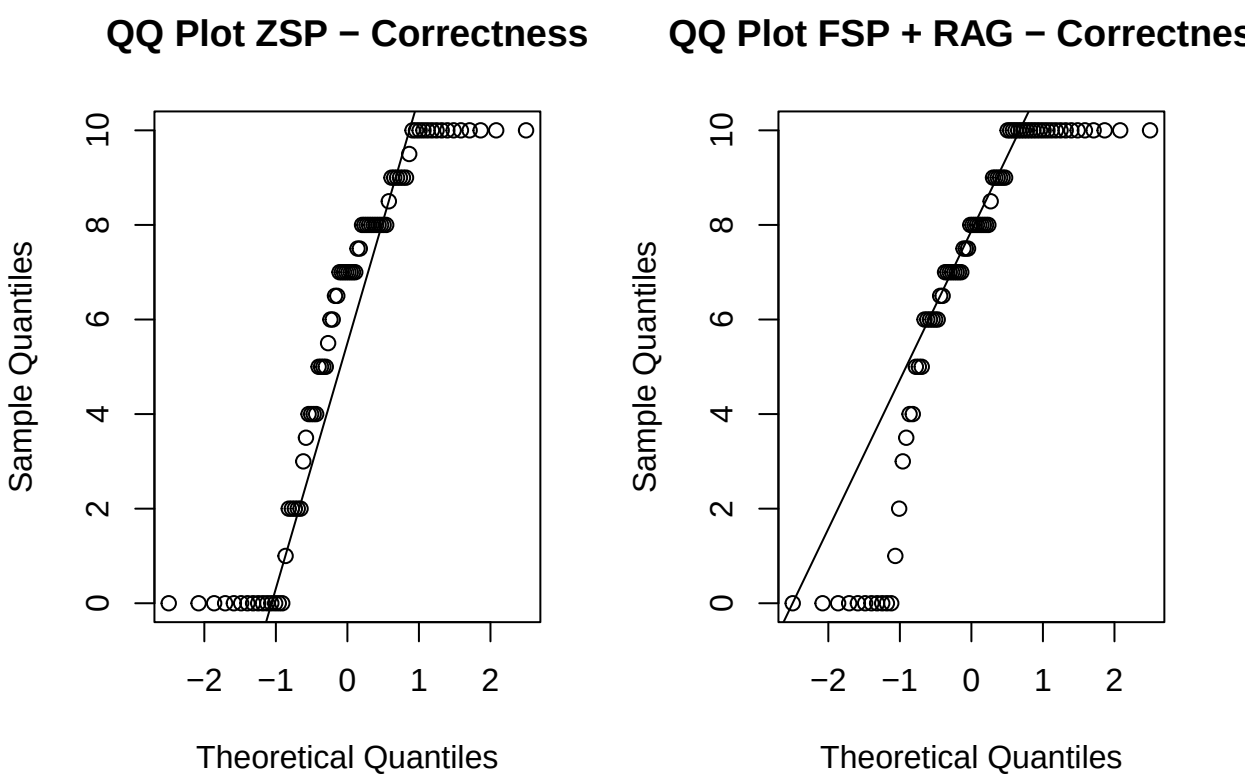
```
df <- read_csv2(".\\LLM_Benchmarks.csv")
```

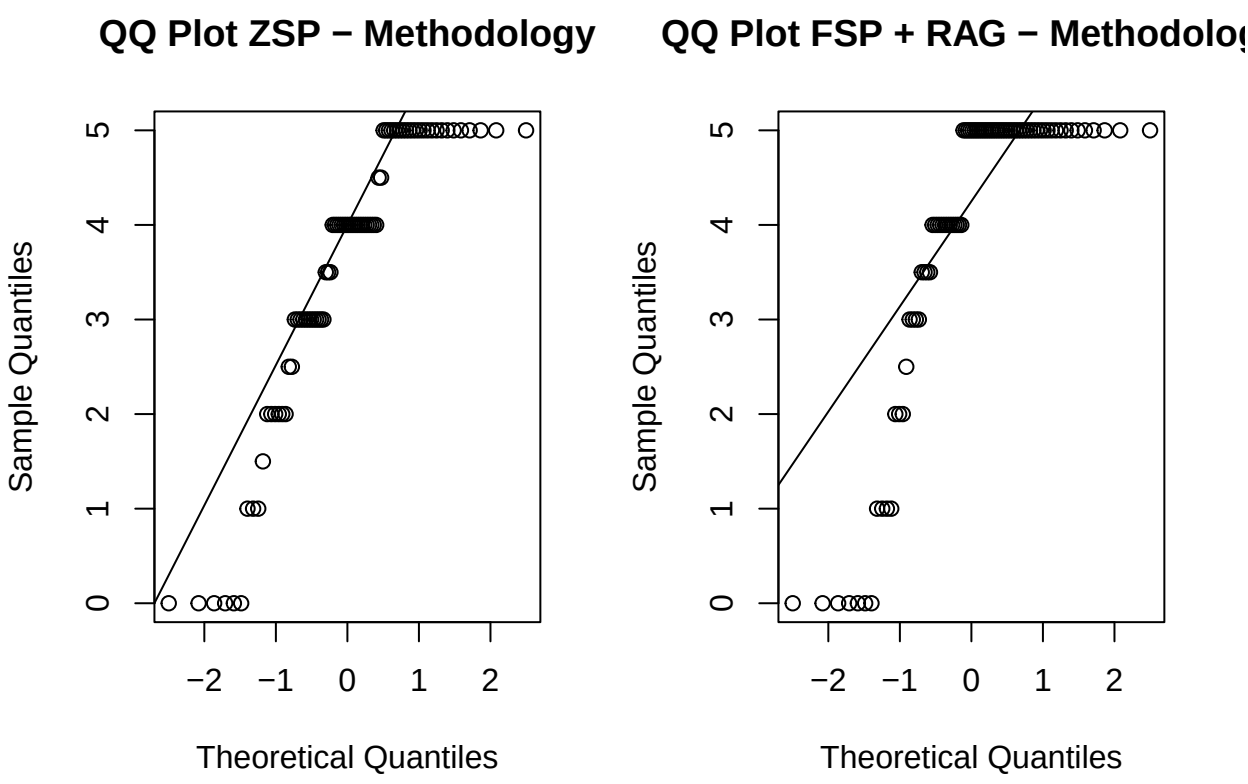
2. Normality Tests (Shapiro-Wilk)

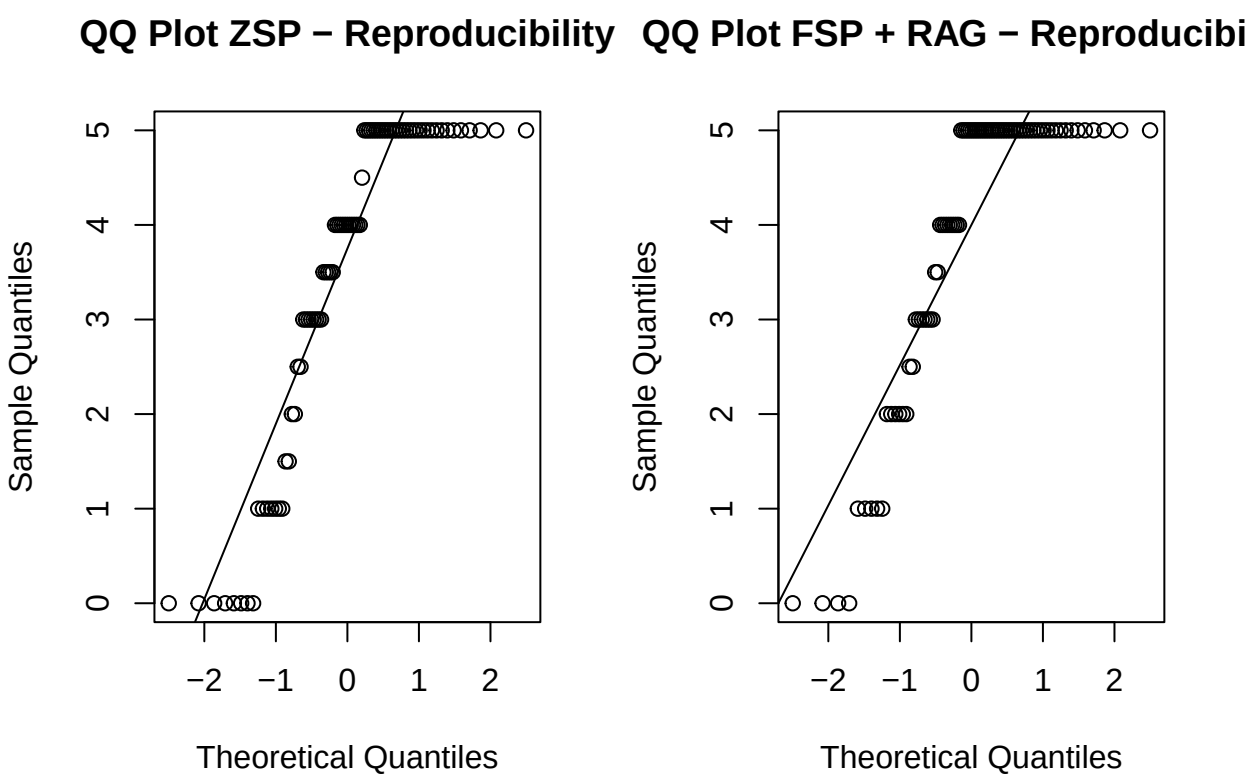
```
shapiro_results <- sapply(c("Correctness", "Methodology", "Reproducibility", "Quality",
 "Score"), function(metric) {
  zsp <- df[[paste0(metric, "_ZSP")]]
  fsp <- df[[paste0(metric, "_FSP_RAG")]]
  p_zsp <- shapiro.test(zsp)$p.value
  p_fsp <- shapiro.test(fsp)$p.value
  c(ZSP_P_VALUE = p_zsp, FSP_RAG_P_VALUE = p_fsp)
})
t(shapiro_results)
```

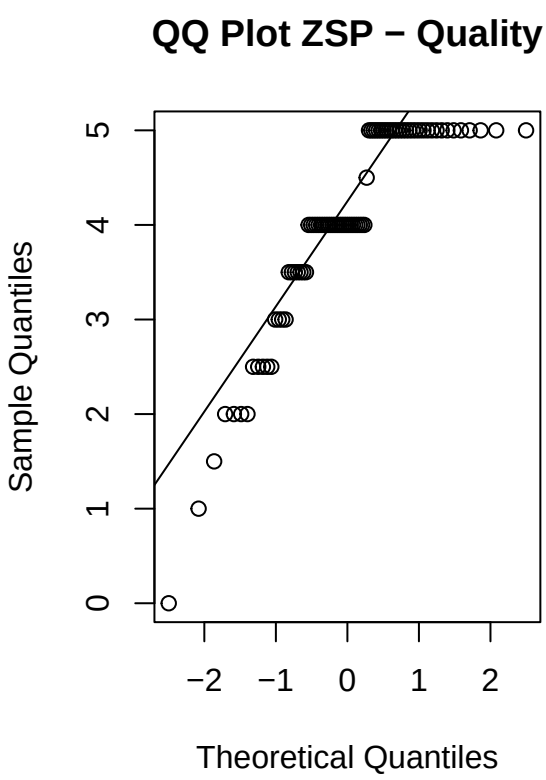
```
##                  ZSP_P_VALUE FSP_RAG_P_VALUE
## Correctness     4.631465e-07    1.545151e-08
## Methodology     1.475666e-07    2.484478e-11
## Reproducibility 8.794292e-09    1.350029e-10
## Quality         3.531027e-08    6.085396e-09
## Score           1.567592e-03    1.538406e-07
```

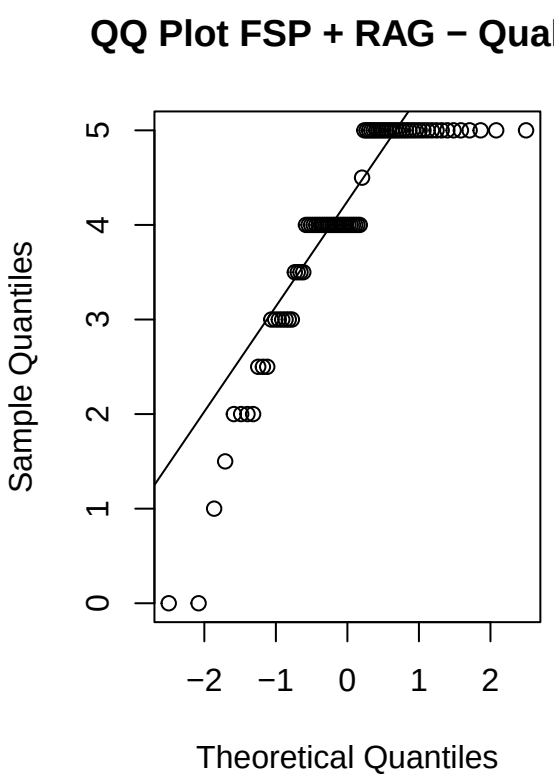
```
# QQ Plots for visualizing normality
metrics <- c("Correctness", "Methodology", "Reproducibility", "Quality", "Score")
par(mfrow = c(1, 2))
for (metric in metrics) {
  zsp <- df[[paste0(metric, "_ZSP")]]
  fsp <- df[[paste0(metric, "_FSP_RAG")]]
  qqnorm(zsp, main = paste("QQ Plot ZSP -", metric)); qqline(zsp)
  qqnorm(fsp, main = paste("QQ Plot FSP + RAG -", metric)); qqline(fsp)
}
```

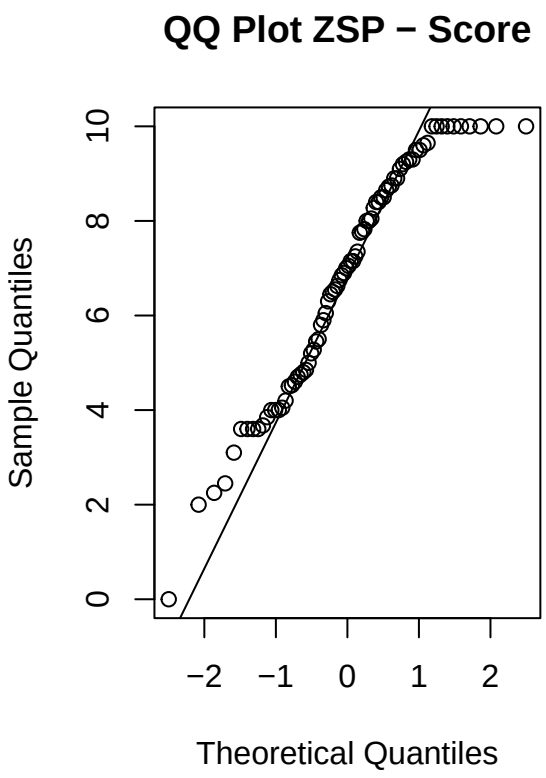

QQ Plot ZSP − Correctness
QQ Plot FSP + RAG − Correctnes
Sample Quantiles
Theoretical Quantiles
0
2
4
6
8
10
−2
−1
0
1
2

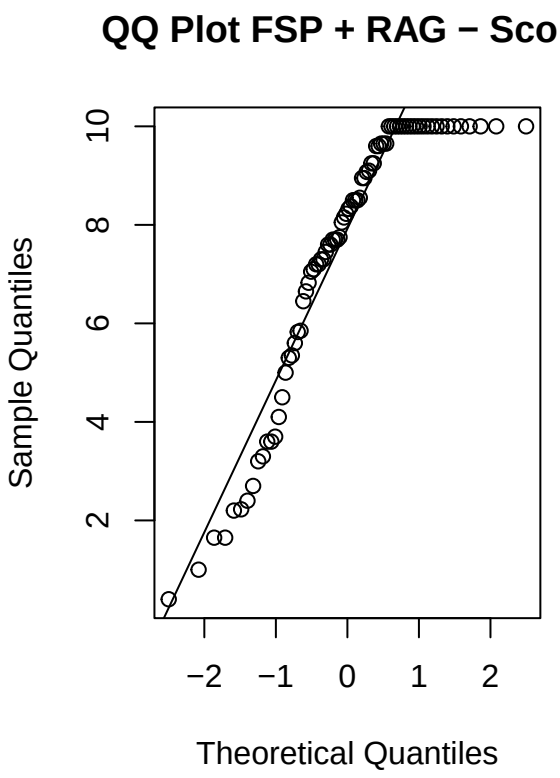

QQ Plot ZSP − Methodology
QQ Plot FSP + RAG − Methodolo
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
5
−2
−1
0
1
2


QQ Plot ZSP − Reproducibility
QQ Plot FSP + RAG − Reproducibi
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
5
−2
−1
0
1
2


QQ Plot ZSP − Quality
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
5
−2
−1
0
1
2


QQ Plot FSP + RAG − Quality
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
5
−2
−1
0
1
2






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

3. Hypothesis Testing (Paired t-test or Wilcoxon)

```
test_results <- lapply(c("Correctness", "Methodology", "Reproducibility", "Quality",
 "Score"), function(metric) {
  zsp <- df[[paste0(metric, "_ZSP")]]
  fsp <- df[[paste0(metric, "_FSP_RAG")]]
  if (shapiro.test(zsp)$p.value > 0.05 && shapiro.test(fsp)$p.value > 0.05) {
    test <- t.test(zsp, fsp, paired = TRUE, alternative = "less")
  } else {
    test <- wilcox.test(zsp, fsp, paired = TRUE, alternative = "less")
  }
  data.frame(Metric = metric, p_value = test$p.value, statistic = test$statistic)
})
do.call(rbind, test_results)
```

```
##             Metric     p_value statistic
## V      Correctness 0.016523477     696.5
## V1     Methodology 0.016999021     458.0
## V2 Reproducibility 0.026197329     497.0
## V3         Quality 0.443124374     673.0
## V4           Score 0.004494911     986.5
```

- H0 (p-value >= 0.05): The mean scores with ZSP are greater than or equal to those with FSP+RAG.
- H1 (p-value < 0.05): The mean scores with ZSP are lower than those with FSP+RAG (meaning FSP+RAG is superior).

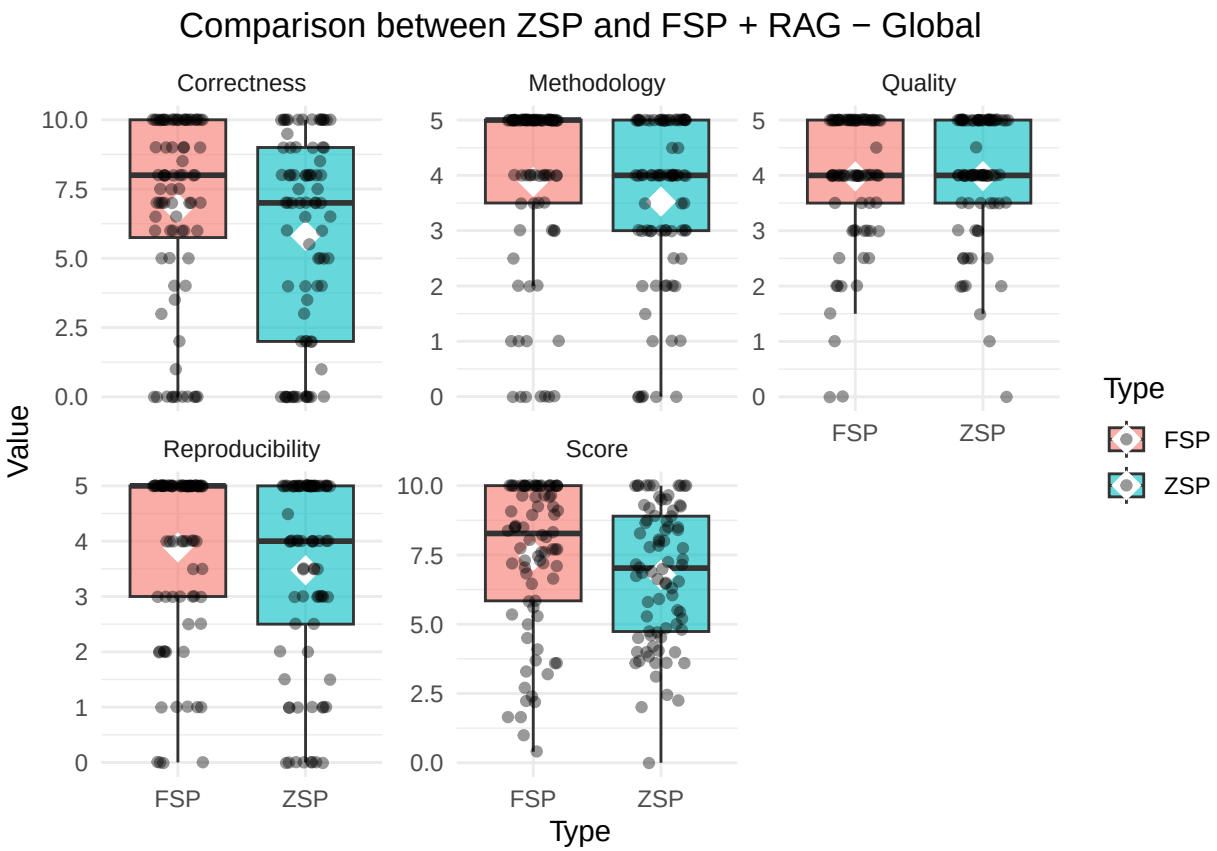
4. Visualization: Comparative Boxplots

```r
# Visualize ranks for Wilcoxon test
ranks_plot_data <- lapply(metrics, function(metric) {
  zsp <- df[[paste0(metric, "_ZSP")]]
  fsp <- df[[paste0(metric, "_FSP_RAG")]]
  if (shapiro.test(zsp)$p.value <= 0.05 || shapiro.test(fsp)$p.value <= 0.05) {
    diff <- zsp - fsp
    signed_ranks <- rank(abs(diff)) * sign(diff)
    data.frame(Subject = seq_along(diff), SignedRank = signed_ranks, Metric = metric)
  }
}) %>% bind_rows()

ggplot(ranks_plot_data, aes(x = factor(Subject), y = SignedRank, fill = Metric)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  facet_wrap(~ Metric, scales = "free_y") +
  theme_minimal() +
  theme(
    axis.text.x = element_blank(),  # Hides x-axis labels
    axis.ticks.x = element_blank(), # Removes x-axis ticks
    plot.title = element_text(hjust = 0.5)
  ) +
  labs(
    title = "Signed Ranks for Wilcoxon Test (ZSP - FSP + RAG)",
    x = NULL,
    y = "Signed Rank"
  )
```


```
df_long <- df %>%
  select(Correctness_ZSP, Correctness_FSP_RAG,
         Methodology_ZSP, Methodology_FSP_RAG,
         Reproducibility_ZSP, Reproducibility_FSP_RAG,
         Quality_ZSP, Quality_FSP_RAG,
         Score_ZSP, Score_FSP_RAG) %>%
  pivot_longer(cols = everything(),
               names_to = c("Metric", "Type"),
               names_sep = "_",
               values_to = "Value")

ggplot(df_long, aes(x = Type, y = Value, fill = Type)) +
  geom_boxplot(alpha = 0.6, outlier.shape = NA) +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 5, color = "white", position = position_d
  geom_jitter(width = 0.2, alpha = 0.4) +
  facet_wrap(~ Metric, scales = "free_y") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Comparison between ZSP and FSP + RAG - Global")
```



5. Analysis by Level

```r
# Grouping by level and calculating the mean per metric
df$Level <- df$Level_ZSP
df_by_level <- df %>%
  group_by(Level) %>%
  summarise(across(ends_with("ZSP"), mean, na.rm = TRUE),
            across(ends_with("FSP_RAG"), mean, na.rm = TRUE))

# Transformation des données pour la visualisation
levels <- sort(unique(df$Level))

get_tests_by_level <- function(niveau) {
  sous_df <- df %>% filter(Level == niveau)
  results <- lapply(c("Correctness", "Methodology", "Reproducibility", "Quality",
   "Score"), function(metric) {
    zsp <- sous_df[[paste0(metric, "_ZSP")]]
    fsp <- sous_df[[paste0(metric, "_FSP_RAG")]]
    if (length(zsp) > 2 && shapiro.test(zsp)$p.value > 0.05 && shapiro.test(fsp)$p.value > 0.05) {
      test <- t.test(zsp, fsp, paired = TRUE, alternative = "less")
    } else {
      test <- wilcox.test(zsp, fsp, paired = TRUE, alternative = "less")
    }
    data.frame(Metric = metric, p_value = test$p.value, statistic = test$statistic)
```

```
  })
  do.call(rbind, results)
}

tests_by_level <- lapply(levels, get_tests_by_level)
names(tests_by_level) <- paste("Level", levels)

tests_by_level
```

```
## $`Level 1`
##             Metric     p_value statistic
## V      Correctness 0.005695738      12.0
## V1     Methodology 0.029648780      10.0
## V2 Reproducibility 0.018399684       7.0
## V3         Quality 0.150401371       7.5
## V4           Score 0.004925565      21.5
##
## $`Level 2`
##             Metric     p_value statistic
## V      Correctness 0.015565517 30.500000
## V1     Methodology 0.001912898  9.500000
## V2 Reproducibility 0.417427485 63.500000
## V3         Quality 0.124106539 34.000000
## t            Score 0.005952462 -2.781041
##
## $`Level 3`
##             Metric   p_value statistic
## V      Correctness 0.4137206      80.0
## V1     Methodology 0.5000000      38.5
## V2 Reproducibility 0.4896625      67.0
## V3         Quality 0.9057033     103.5
## V4           Score 0.4347439     100.0
##
## $`Level 4`
##             Metric    p_value  statistic
## V      Correctness 0.57474540 55.0000000
## V1     Methodology 0.63372787 65.5000000
## V2 Reproducibility 0.03397186 12.0000000
## V3         Quality 0.52519565 53.0000000
## t            Score 0.33596402 -0.4301471
```

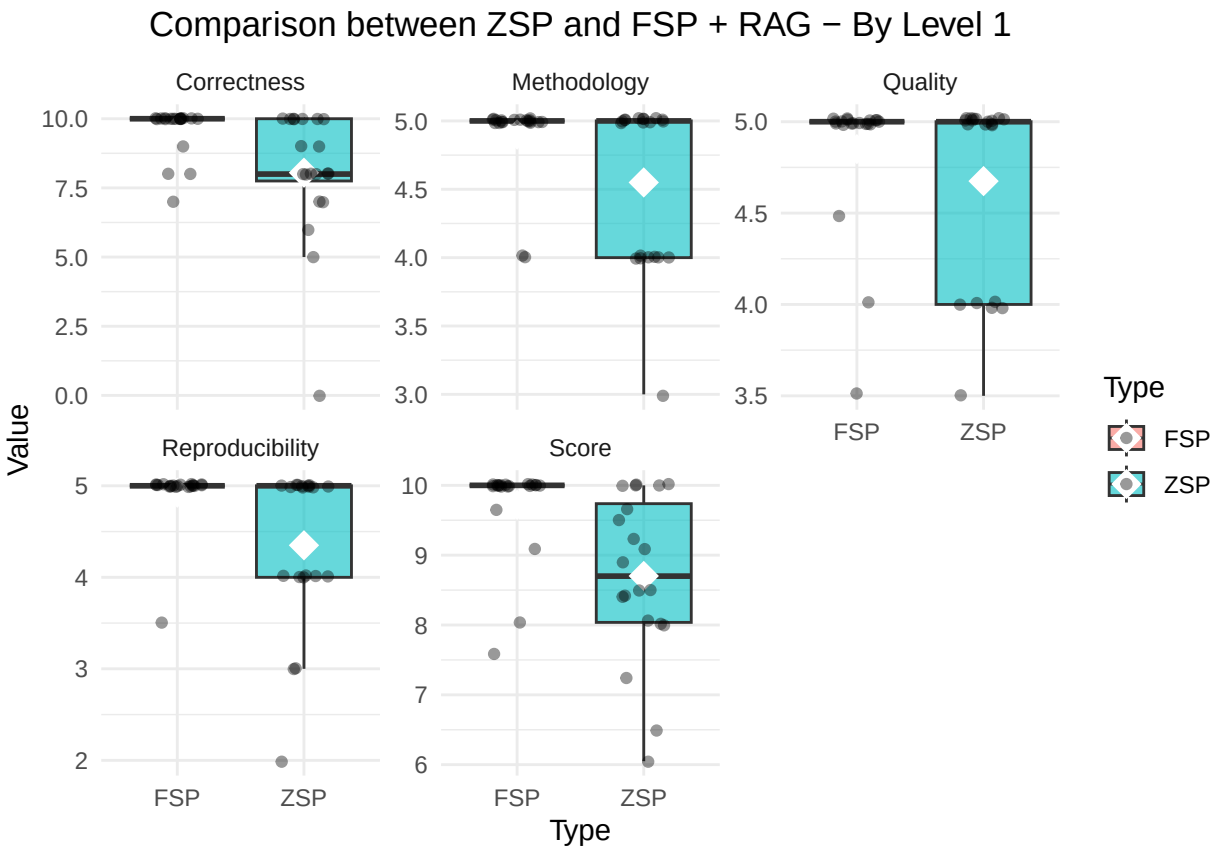
- H0 (p-value >= 0.05): The mean scores with ZSP are greater than or equal to those with FSP+RAG.
- H1 (p-value < 0.05): The mean scores with ZSP are lower than those with FSP+RAG (meaning FSP+RAG is superior).

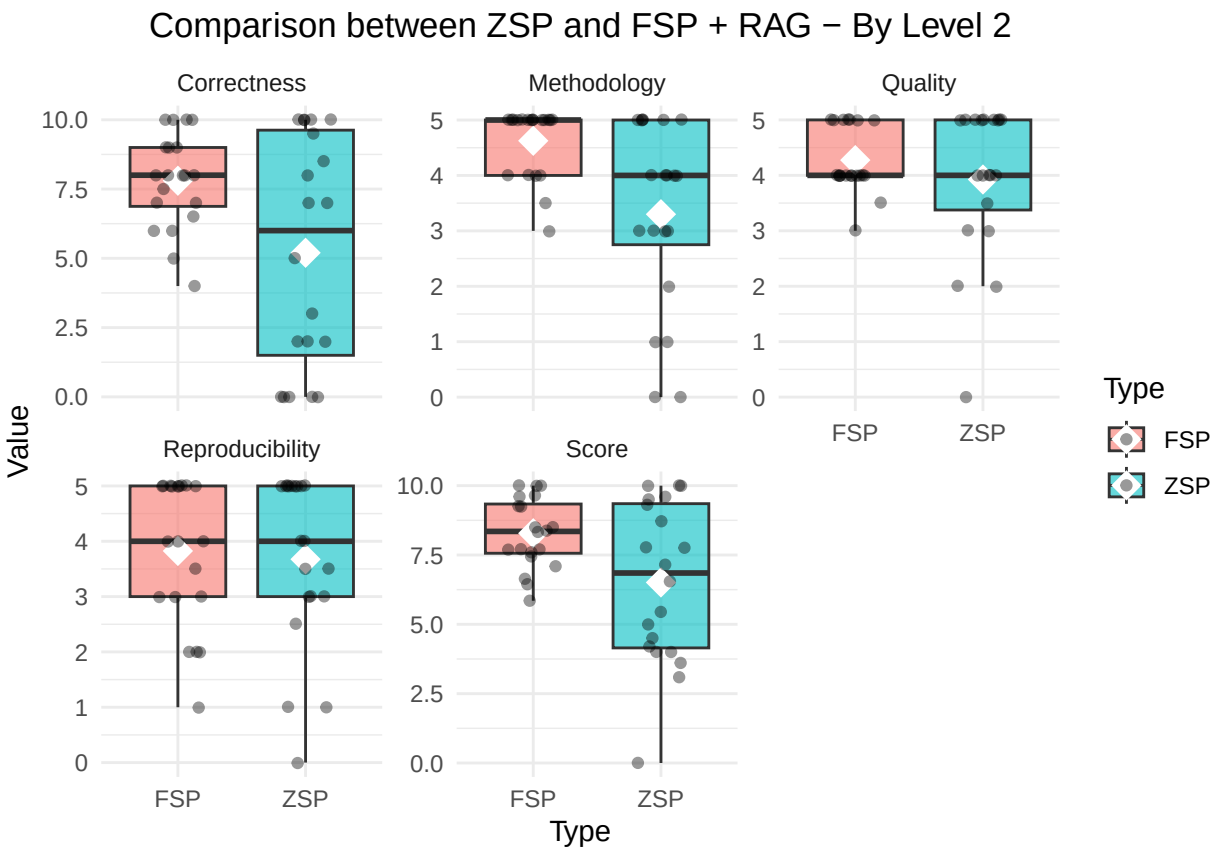
6. Visualizations by Level

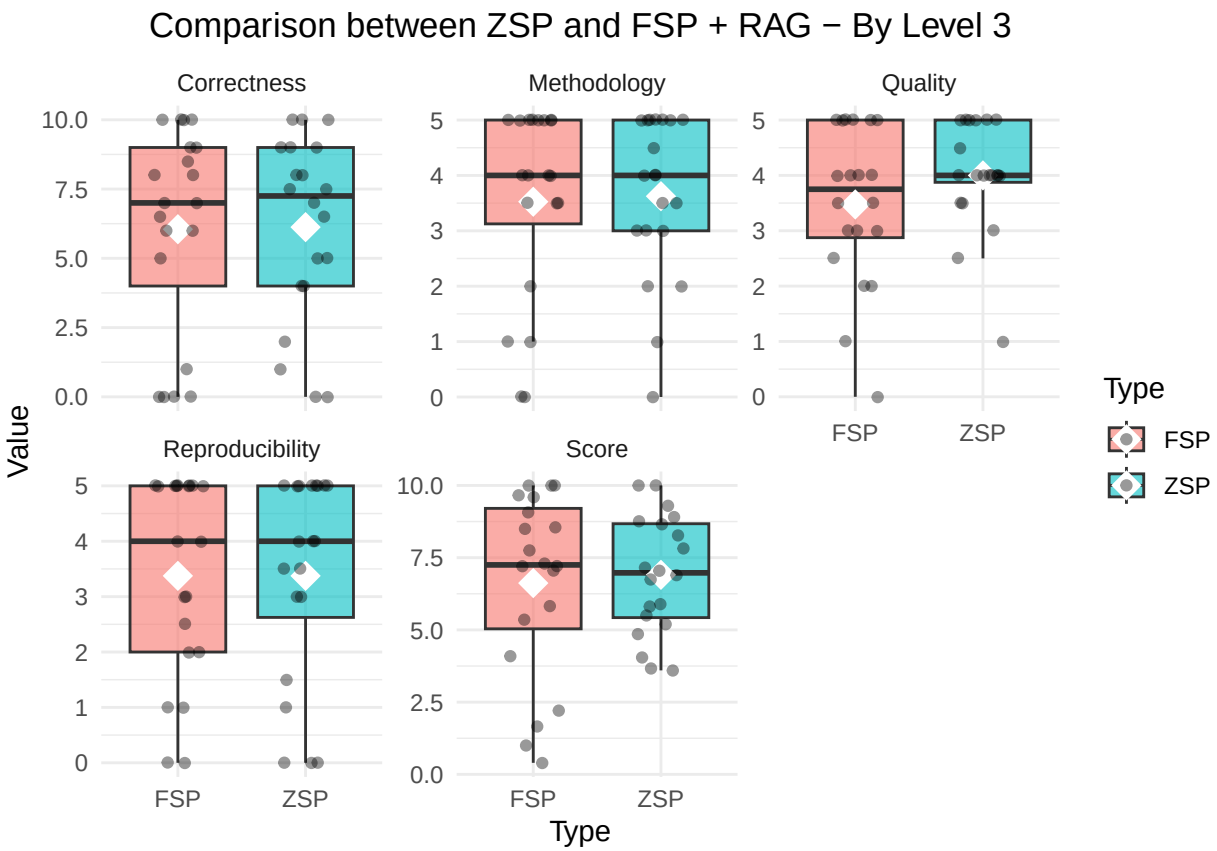
```
for (level in levels) {
  sous_df <- df %>%
    filter(Level == level) %>%
```

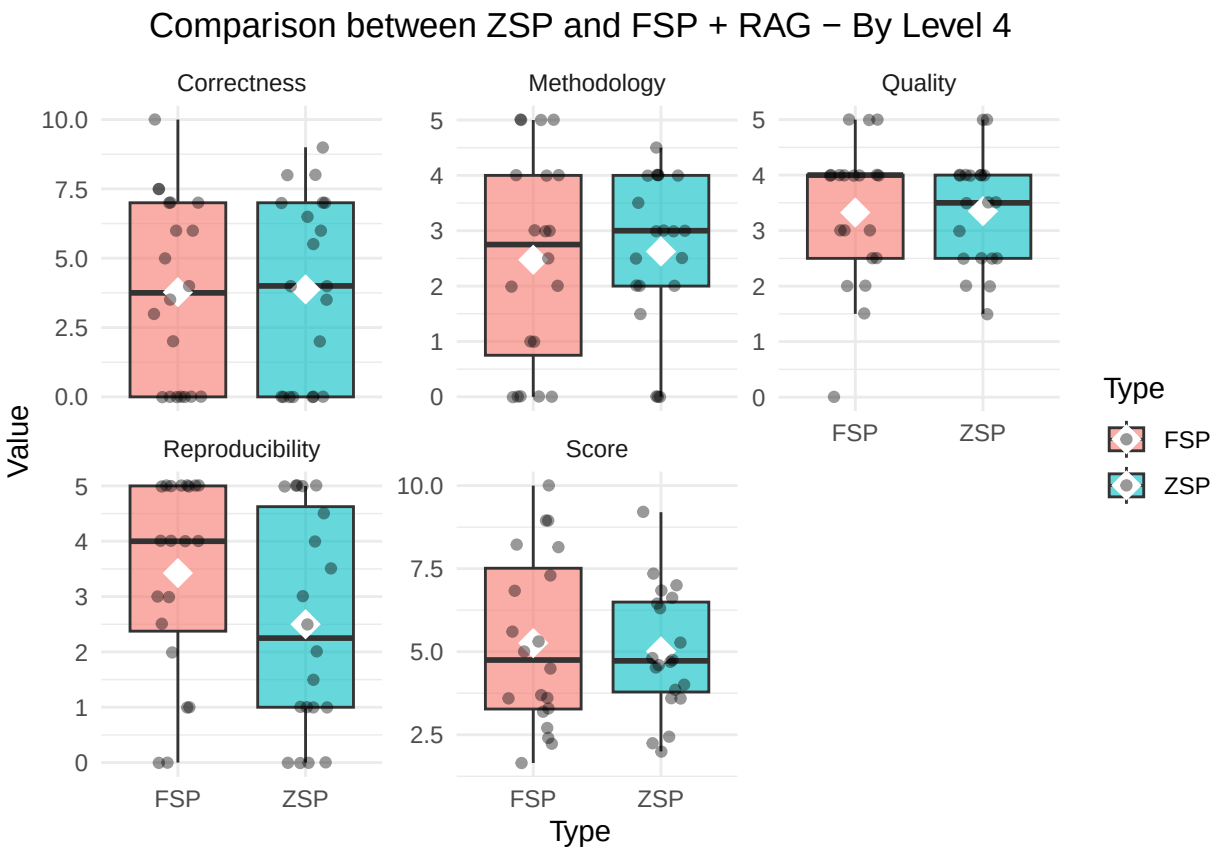
```
    select(Correctness_ZSP, Correctness_FSP_RAG,
           Methodology_ZSP, Methodology_FSP_RAG,
           Reproducibility_ZSP, Reproducibility_FSP_RAG,
           Quality_ZSP, Quality_FSP_RAG,
           Score_ZSP, Score_FSP_RAG) %>%
    pivot_longer(cols = everything(),
                 names_to = c("Metric", "Type"),
                 names_sep = "_",
                 values_to = "Value")

  print(
    ggplot(sous_df, aes(x = Type, y = Value, fill = Type)) +
      geom_boxplot(alpha = 0.6, outlier.shape = NA) +
      stat_summary(fun = mean, geom = "point", shape = 18, size = 5, color = "white", position = positi
      geom_jitter(width = 0.2, alpha = 0.4) +
      facet_wrap(~ Metric, scales = "free_y") +
      theme_minimal() +
      theme(plot.title = element_text(hjust = 0.5)) +
      labs(title = paste("Comparison between ZSP and FSP + RAG - By Level", level))
  )
}
```




Comparison between ZSP and FSP + RAG − By Level 2
Correctness
Methodology
Quality
Reproducibility
Score
Value
Type
FSP
ZSP
10.0
7.5
5.0
2.5
0.0
5
4
3
2
1
0


Comparison between ZSP and FSP + RAG − By Level 3
Correctness
Methodology
Quality
Reproducibility
Score
Value
Type
FSP
ZSP
10.0
7.5
5.0
2.5
0.0
5
4
3
2
1
0


Comparison between ZSP and FSP + RAG − By Level 4
Correctness
Methodology
Quality
Reproducibility
Score
Value
Type
FSP
ZSP