Compute Interrater/Intercoder Reliability

Between two individual Coders

Cohen's Kappa

- Represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the raters.
- >0.6 = **moderate agreement**

```
kapp = cohen_kappa_score(coder_1["label"], coder_2["label"])
print(f"kappa: {kapp:.2f}")
```

```
kappa: 0.68
```

Percent Agreement

```
print(f"percent agreement: {accuracy_score(coder_1['label'], coder_2['label'])*100:.2f}")
```

```
percent agreement: 89.22
```

Category Specific Agreement

From the agreement package: "Specific agreement is an index of the reliability of categorical measurements. It describes the amount of agreement observed with regard to each possible category. With two raters, the interpretation of specific agreement for any category is the probability of one rater assigning an item to that category given that the other rater has also assigned that item to that category. With more than two raters, the interpretation becomes the probability of a randomly chosen rater assigning an item to that category given that another randomly chosen rater has also assigned that item to that category. When applied to

binary (i.e., dichotomous) data, specific agreement on the positive category is often referred to as positive agreement (PA) and specific agreement on the negative category is often referred to as negative agreement (NA). In this case, PA is equal to the F1 score frequently used in computer science." "

```
library(agreement)
data = read.csv("data/awo_coders_long.csv")
results <- cat_specific(data)
summary(results, ci = TRUE, type = "perc")
```

```
Call:
cat_specific(.data = data)

Objects =   306
Raters =    2
Categories =    {0, 1}

Category-Specific Agreement with Bootstrapped CIs

    Estimate   lower   upper
0      0.931   0.906   0.953
1      0.752   0.662   0.827
```

Between two individual coders and a labeling version that multiple coders worked on in intervals

Fleiss's Kappa

- See Wikipedia
- The measure calculates the degree of agreement in classification over that which would be expected by chance.
- >0.6 = **Moderate Agreement**

```
arr, cats = aggregate_raters(all_coders)
kapp = fleiss_kappa(arr)
print(f"kappa: {kapp:.2f}")
```

```
kappa: 0.68
```

Percent Agreement

```
percent_agreement = acc_df["agreement_pct"].mean()
print(f"Percent agreement: {percent_agreement}")
```

```
Percent agreement: 88.4525816993464
```

Category Specific Agreement

```
library(agreement)
data = read.csv("data/all_coders_long.csv")
results <- cat_specific(data)
summary(results, ci = TRUE, type = "perc")
```

```
Call:
cat_specific(.data = data)

Objects =   306
Raters =    3
Categories =    {0, 1}

Category-Specific Agreement with Bootstrapped CIs

    Estimate   lower   upper
0      0.924   0.903   0.943
1      0.757   0.690   0.812
```

Conclusion

According to McHugh (2012): "Perhaps the best advice for researchers is to calculate both percent agreement and kappa. If there is likely to be much guessing among the raters, it may make sense to use the kappa statistic, but if raters are well trained and little guessing is likely to exist, the researcher may safely rely on percent agreement to determine interrater reliability."

AWO coders were trained and did not guess, so we rely on percent agreement. Therefore, interrater/intercoder reliability should is sufficient with strong agreement overall. For the specific labels, there strong agreement for the negative label 0 and moderate agreement for the positive label.