**NOTE: I have used ChatGPT extensively for this assignment.**

**R1. Reliability of all the observed data(that is, 3 observed columns):**

Reliability is based on observed scores, not estimated ones. Reliability measures consistency across a set of observed values where 0 means unreliable and 1 means highly reliable. Reliability score is given by this formula(Simplified Coefficient of Variation-based approach): **R = 1 – (mean/standard deviation**). Mean and standard deviation for each row(where each row has 3 values) is calculated and used to calculate the reliability score.

A screenshot of a computer

Description automatically generated

Reliability Score

P1 0.93

P2 0.86

P3 0.89

P4 0.80

P5 0.92

P6 0.82

P7 0.94

P8 0.94

P9 0.87

P10 0.82

P11 0.89

P12 0.86

As we can see, 8 of the values are between 0.8 to 0.89 and 4 of the values are >= 0.9. This points to the data being reliably good(referencing the table above) but not reliably excellent(for the most part).

Code: <https://colab.research.google.com/drive/1_gzJoDRPWEY29I8MOUvMXfNucZzbA3UV?usp=sharing>

(**if the csv file is deleted by the time you check this code—colab deletes files for runtime purposes, please upload a csv file of the problem set – this applies to every link I share from now on**)

**R2. Reliability of most of the observed data(that is, 2 observed columns, which means we exclude the caregiver observed score):**

Reliability is calculated similar to before but in this case, we remove the caregiver observed score. Why do we do that? For the doctor and the therapist scores, we know that it’s **one** doctor and one therapist giving the scores. For the caregiver scores, we know that it’s **12** caregivers giving 12 scores, which introduces a lot of variability to the data, one which is not desirable. Hypothetically, if it was **one** caregiver, we would have 3 people giving scores. In the current data, we have 14 people giving scores. One person can be very error prone, 14 people would be even more error prone. So, we can say that the reliability scores given by the caregivers may introduce more noise than necessary. Hence, we remove the caregiver score altogether.

Reliability Score

P1 1.00

P2 0.92

P3 0.89

P4 0.89

P5 1.00

P6 1.00

P7 0.93

P8 0.93

P9 1.00

P10 1.00

P11 0.89

P12 0.86

As we can see, the situation has reversed now. There are 4 values between 0.8 and 0.89 and 8 values >= 0.9. Of note, on a 1-to-1 comparison with the previous reliability scores from R1, 7 rows increased in reliability (P1, P2, P4, P5, P6, P9, P10), 3 rows did not change(P3, P11, P12), and 2 rows decreased(P7, P8). Overall, the data is more reliable if we exclude caregivers from the calculation. And finally, the data is reliably excellent, for the most part.

Code:

<https://colab.research.google.com/drive/1cHmTgpuXAgCSvtvXPpKiJutW7laKoAi9?usp=sharing>

**V. Validity plots for various combinations.**

Validity approximates how close estimated scores are to observed scores. In the case of perfect matches, validity is 0(imagine observed is 10 and estimated is 10) and in the worst case, validity is upper – lower bound of data. In our case, upper bound of FRS is 5 and lower bound is 1, therefore, worst case validity is 4. Best case validity is 0. We will use Mean Absolute Error(MAE) to calculate validity. MAE = 1/n \* Summation|(yestimated - yobserved)|.

Please note that Google Colab does not allow many sessions to run. Therefore, I cannot link my code to all of the 8 scenarios I want to run. However, I will link my code for some of the scenarios(**V1 and V8**) and as it is, the codes are very similar across the 8 scenarios.

**V1: MAE(mean of 3 observed vs estimated in-clinic)**

Code: <https://colab.research.google.com/drive/1Hjc7l9hsyyitIXZ_8bAEZTYlvldpYKa3?usp=sharing>

**MAE(Validity Score) = 0.38**

**A graph with blue and orange lines

Description automatically generated**

**V2: MAE(mean of 3 observed vs mean of estimated 2 remotes)**

**MAE(Validity Score) = 0.46**

**A graph with blue and orange lines

Description automatically generated**

**V3: MAE(mean of 2 observed without caregiver vs estimated in-clinic)**

**MAE(Validity Score) = 0.29**

A graph with blue and orange lines

Description automatically generated

**V4: MAE(mean of 2 observed without caregiver vs mean of estimated 2 remotes)**

**MAE(Validity Score) = 0.31**

**A graph with blue and orange lines

Description automatically generated**

**V5: MAE(Caregiver Observed vs estimated in-clinic)**

**MAE(Validity Score) = 0.62**

**A graph with blue and orange lines

Description automatically generated**

**V6: MAE(Caregiver Observed vs mean of estimated 2 remotes)**

**MAE(Validity Score) = 0.75**

A graph with blue and orange lines

Description automatically generated

**V7: MAE(median of 3 observed vs estimated in-clinic)**

**MAE(Validity Score) = 0.35**

**A graph with blue and orange lines

Description automatically generated**

**V8: MAE(median of 3 observed vs mean of estimated 2 remotes)**

Code: <https://colab.research.google.com/drive/1LI0sznEcn7QmOxZeKf2R0OV3LKMOwyh9?usp=sharing>

**MAE(Validity Score) = 0.46**

**A graph with blue and orange lines

Description automatically generated**

**Final Notes on validity:**

As you can see, median and mean of 2 and 1 values is literally the same, so doing analysis for median of 2 observed values without caregiver and median of 1 caregiver value is rendered moot.

From these 8 scenarios, we have these values:

1. **Validity(mean of 3 observed vs estimated in-clinic) = 0.38**
2. **Validity(mean of 3 observed vs mean of estimated 2 remotes) = 0.46**
3. **Validity(mean of 2 observed without caregiver vs estimated in-clinic) = 0.29**
4. **Validity(mean of 2 observed without caregiver vs mean of estimated 2 remotes) = 0.31**
5. **Validity(Caregiver Observed vs estimated in-clinic) = 0.62**
6. **Validity(Caregiver Observed vs mean of estimated 2 remotes) = 0.75**
7. **Validity(median of 3 observed vs estimated in-clinic) = 0.35**
8. **Validity(median of 3 observed vs mean of estimated 2 remotes) = 0.46**

Following the ideas of reliability, it made sense to compare validity with and without caregiver observed scores and validity with caregiver observed scores by itself. And we can see that caregiver scores by itself reduce the validity of the data(scenarios 5 and 6). I haven’t made plots for doctor by itself or therapist by itself vs ground truth scores(partly because we would have too many plots) but a quick glance through the data reveals that the scores from therapist and doctor are closer to estimated scores than caregiver data. Additionally, when we compare validity of 1 therapist and 1 doctor vs estimated scores(scenarios 3 and 4), we can see that validity is really good. Recall that validity near 0 means really valid and validity near 4 means almost invalid. Scenarios 7 and 8 are copies of scenarios 1 and 2, the only metric that is changed is that we are using medians. Finally, scenarios 1,2,7, and 8 reveal that inclusion of caregiver scores bring down the validity of the data. Given that the range of validity is 0-4, and all the validity scores are less than 1, I would say that the data is highly valid. If the scores started approaching 2, I would begin to doubt the validity but that’s not the case. As a note, for future calculations, I would probably have 1 caregiver as opposed to 12(as explained in the reliability section earlier) because I think that is bringing down the reliability and validity overall.

MAE(caregiver vs estimated in-clinic)

MAE(caregiver vs mean of estimated 2 remotes)

Next 6 are mirrors of the 6 above but we calculate median within the brackets.