The goal of the Project is to estimate a risk score for 499 files based on a sample of 81 files. The total risk score will be calculated based on a weighted average using pre-determined risk scores for 6 different factors: Size of Case, Scheme, Age, Age to Drawdown, Investment profile and Investment risk. Each factor will be divided in 10 levels and specific scores will be given to each case, depending on the level it falls. The weights were created so that the maximum score given to a file is 100. Three factors were given a weight of 10 (Size of Case, Scheme and Age), two factors a weight of 20 (Age to Drawdown and Investment Risk) and one factor was given a weight of 30 (Investment Profile). For instance, if a case falls on the highest risk level for Investment profile (i.e. having 60% of equity or more) it will be given a score of 30, which will be subsequently summed with the scores of each additional factor, summing up to a maximum of 100. The detailed values for each factor's level can be found attached.

Commented [RADR1]: I suggest you add the table you shared in the PDF to an addendum/appendix.

A probabilistic sampling strategy will be employed, and 81 cases will be randomly selected from a total population of 499 files (n = 81). A stratified random sampling will not be used since there are no known levels of disaggregation that are relevant to determining the risk scores apart from the ones already used in the calculation. Thus, a simple random sampling will be employed to estimate the population's risk score.

Since the mean and standard deviation of the population are unknown, the confidence interval around the population's mean will be calculated as follows:

$$\mu = \bar{x} \pm t \times \frac{s}{\sqrt{n}}$$

Where,

μ = mean risk score for the population

$\bar{x}$ = mean risk score for the sample

t = critical value of a t-distribution relative to 95% confidence level.

s = standard deviation of the risk score for the sample

n = sample size (81).

The critical value used in the formula will be 1.99 which refers to the value relative to a t-distribution with 80 degrees of freedom (n-1). The calculated confidence interval for the mean will mean that the true risk score for the whole population of 499 files will be between the two limits of the interval with 95% statistical confidence.