Justification for Pooled Standard Deviation Formulae

Consider two groups of data:

group	mean	st. dev	n
1	9.7	2.5	50
2	17.3	6.8	200.

if you pooled them, to calculate the mean of all 250 pieces of data, you would do:

$$\bar{x}_{POOL} = \frac{50\bar{x}_1 + 200\bar{x}_2}{250}$$

so it follows that to get the pooled standard deviation, you would do

$$\sigma^2_{POOL} = \frac{50\sigma_1^2 + 200\sigma_2^2}{250}$$

However, as we are using sample standard deviations to estimate population standard deviations, recall the formulae for $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$ or $s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$

and $\sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}$ or $\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$

Multiplying each formula by the denominator, gives

$$n\sigma^2 = (n-1)s^2$$

This leads to $50\sigma_1^2 = 49s_1^2$ and $200\sigma_2^2 = 199s_2^2$

and we'd adjust the ÷ by 250 to be ÷(49+199), which is (÷248)

so $$s^2_{POOL} = \frac{49s_1^2 + 199s_2^2}{248}$$

or, more generally, $$s^2_{POOL} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

$$\Rightarrow \quad s^2_{POOL} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$