Justification of Formula for Determining Minimum Rank Sum in Mann Whitney Test

In the SQA AH Statistics course, we conduct a Mann-Whitney test for non-paired data with two samples, with sample sizes m and n, where $m \le n$. We calculate the rank sum of the group of size m, and compare its value to the critical values listed in Table 8 of the AH Statistics 'Statistical Tables and Formulae' booklet. However, the value of the rank sum is dependent upon whether the data is ranked from 1 to $(m+n)$ taking the lowest value to be rank 1, and highest value to be rank $(m+n)$ **or** whether rank 1 is the highest value and rank $(m+n)$ is the lowest value.

Numerical Example.

Consider two samples A and B, of sizes m = 2 and n = 3 respectively, ranked from lowest to highest value:

sample	A	B	A	B	B
Rank (low to high)	1	2	3	4	5

giving W_A = 1 + 3 = 4

Now, if the same data had been ranked in the reverse order (from highest to lowest) it would have looked like this:

sample	A	B	A	B	B
Rank (high to low)	5	4	3	2	1

giving W_A = 5 + 3 = 8

The tables in the SQA AH Statistics booklet are set up to use the **smallest** rank sum from these two possible ranking choices.
So, if you had worked out W_A = 8, you would have to also check whether ranking in the reverse order would give you a lower rank sum. This can either be done by manually re-ranking the data (as done above) and checking, or using the formula stated in the Data Booklet of $m\left(m+n+1\right)-W_A$.

Therefore, the formula is just a shortcut to the manual re-ranking process, and in our numerical example it gives 2(2+3+1)—W_A = 12—8 = 4. Notice that when the ranks were reversed from 'low-to-high', to 'high-to-low', each new rank = 6 — old rank. This 6 came from $m+n+1$.

General Proof

Consider a group of m values ranked in order, from 'low-to-high'.
Let these rank values be $r_1, r_2, \dots, r_{m-1}, r_m$

So, $W_m = \sum_{i=1}^{m} r_i$

If these values are then ranked in reverse order from 'high-to-low', each r_i now becomes $(m+n+1)-r_i$
Then,

$$\begin{aligned} W_m &= \sum_{i=1}^{m}\left[(m+n+1)-r_i\right] \\ &= \sum_{i=1}^{m}(m+n+1)-\sum_{i=1}^{m} r_i \\ &= (m+n+1)\sum_{i=1}^{m}1-\sum_{i=1}^{m} r_i \\ &= m(m+n+1)-\sum_{i=1}^{m} r_i \end{aligned}$$

And so, the value of W that is used is the minimum of W_m and $m\left(m+n+1\right)-W_m$.