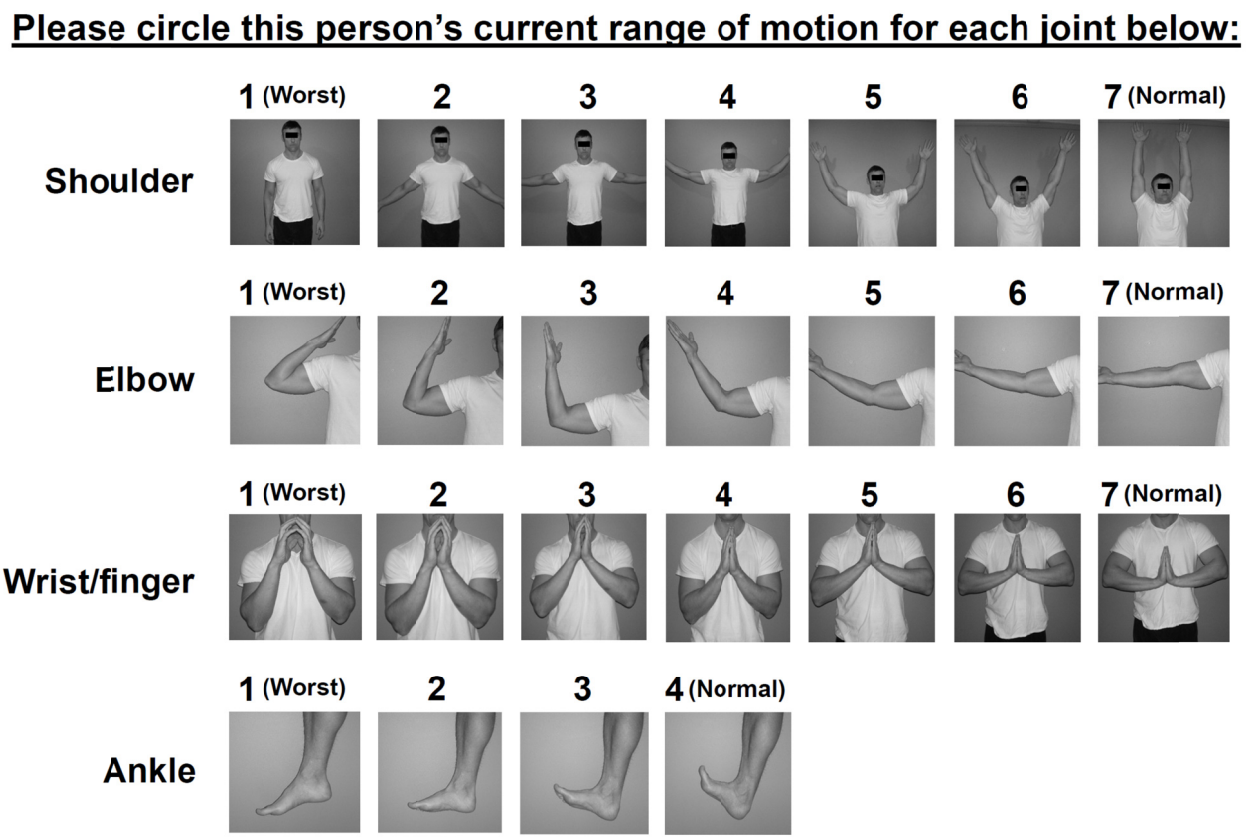
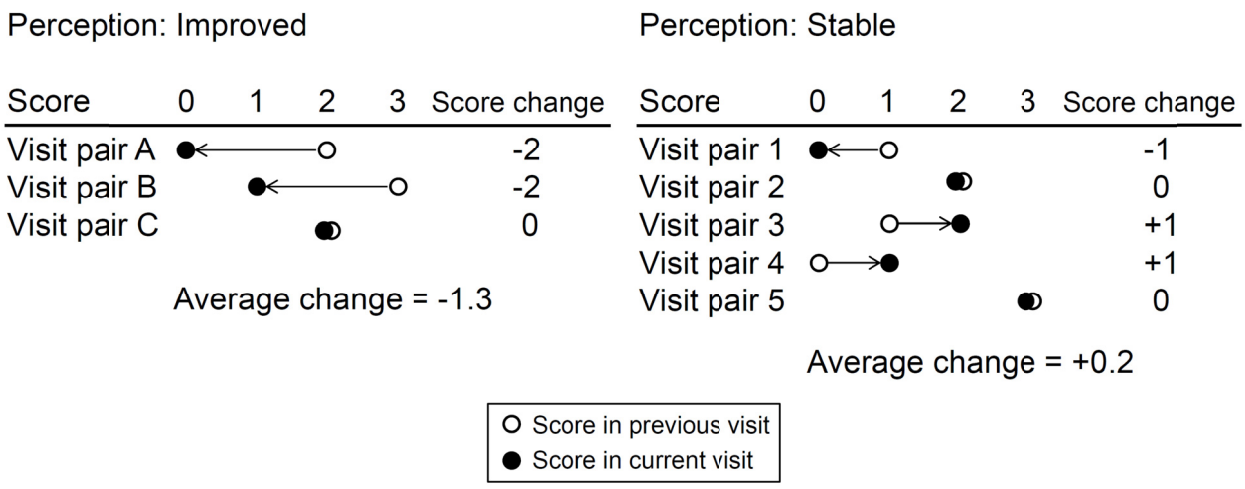
Supplementary Figure S1. Photographic Range of Motion (P-ROM) scale. Lower scores indicate more limited range of motion. The P-ROM total score is the sum of scores in all 4 joints, and ranges 4-25 points.



Supplementary Figure S2. An example of calculating the "estimated difference" between improvement and stability for the NIH scale. NIH scores in previous visits are shown as open symbols, and NIH scores in the current visits are shown as filled symbols. Directions of arrow indicate directions of the change in the NIH score between visit pairs. In this example (left panel), 3 paired visits that were perceived as joint improvement are averaged. Similarly (right panel), 5 paired visits that were perceived as joint stability are averaged. The average change in NIH score for all visits associated with perceived improvement is ([-2] + [-2] + 0)/3 = -1.3. Similarly, the average change in NIH score for all visits associated with perceived stability is ([-1] + 0 + 1 + 1 + 0)/5 = +0.2. Therefore, the estimated difference in these average change scores for the improved group compared to the stable group is [-1.3] – [+0.2] = -1.5. This final derived "estimated difference" represents the difference between the average change scores for the comparison group (improved) relative to the reference group (stable). This concept was modeled in the mixed linear models and the point estimate and 95% confidence interval were calculated. The same method was used for estimating the difference for the group of visits associated with perceived worsening compared to perceived stability.



The "absolute estimated difference" for the group of visits with perceived improvement compared to the group of visits with perceived stability is 1.5.