**Additional file 4. Quality assessment for studies using a cohort and before-and-after study design (*n* = 6).**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sample** | **Study design** | **Measurement** | **Statistical Analysis** | **Total Points1** | **Score** | **Quality** |
| **First Author, Year** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** |
| Witjes, 2019 | 0 | 0 | 1 | N/A | 1 | N/A | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 8/16 | 0.50 | Weak |
| Dominguez-Gil, 2017 | 0 | 0 | 1 | N/A | 1 | N/A | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 6/16 | 0.44 | Weak |
| Vorstius Kruijff, 2016 | 0 | 0 | 1 | N/A | 1 | N/A | 1 | 0 | 1 | 1 | 0 | 0 | N/A | 5/15 | 0.33 | Weak |
| Czerwinski, 2014 | 0 | 0 | 1 | 1 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 6/18 | 0.33 | Weak |
| Sikora, 2014 | 0 | 0 | 1 | N/A | 0 | N/A | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2/16 | 0.13 | Weak |
| Santiago, 2005 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 4/18 | 0.22 | Weak |

*Note.* Total Points: No. of points possible = (18 – N/A)

Key: < 0.50= weak; 0.51-0.65= low-moderate; 0.66-0.79= high-moderate; > 0.80= strong

DV = dependent variable; N/A = Not Applicable

Assessment:

1. Was probability sampling used?

2. Was sample size justified to obtain an appropriate power?

3. Are the individuals selected to participate in the study likely to be representative of the target population?

4. 1) One pre-test or baseline and several post-test measures; 2) simple before-and-after study; or 3) Post-test only?

5. Does the study employ a comparison strategy?

6. Group comparisons were same for all occasions

7. Was the DV 1) directly measured (i.e., observed) or taken from an administrative database or chart or 2) self-reported?

8. Was the DV reliably and validly measured? That is, measured using an instrument with reported (previously or for this study) 1) reliability indices and 2) validity assessments?

9. Was (were) the statistical test (s) used appropriate for the main outcome?

10. Were p values reported?

11. Were confidence intervals reported?

12. Were missing data managed appropriately?

13. Is attrition rate < 30%?