Validation of the English language Forgotten Joint Score-12 as an outcome measure for total hip and knee arthroplasty in a British population

Citation for published version:

Digital Object Identifier (DOI):
10.1302/0301-620X.99B2.BJJ-2016-0606.R1

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Bone and Joint Journal

Publisher Rights Statement:
This is the author's final peer-reviewed manuscript as accepted for publication.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Validation of the English language Forgotten Joint Score-12 as an outcome measure for total hip and knee arthroplasty in a British population

Aims
To validate the English language Forgotten Joint Score-12 (FJS-12) as a tool to evaluate the outcome of hip and knee arthroplasty in a United Kingdom population.

Patients and Methods
All patients undergoing surgery between January and August 2014 were eligible for inclusion. Prospective data were collected from 205 patients undergoing total hip arthroplasty (THA) and 231 patients undergoing total knee arthroplasty (TKA). Outcomes were assessed with the FJS-12 and the Oxford Hip and Knee Scores (OHS, OKS) pre-operatively, then at six and 12 months post-operatively. Internal consistency, convergent validity, effect size, relative validity and ceiling effects were determined.

Results
Data for the TKA and THA patients showed high internal consistency for the FJS-12 (Cronbach $\alpha = 0.97$ in TKAs, 0.98 in THAs). Convergent validity with the Oxford Scores was high ($r = 0.85$ in TKAs, $r = 0.79$ for THAs). From six to 12 months, the change was higher for the FJS-12 than for the OHS in THA patients (effect size $d = 0.21$ versus $-0.03$). Ceiling effects at one-year follow-up were low for the FJS-12 with just 3.9% (TKA) and 8.8% (THA) of patients achieving the best possible score.

Conclusion
The FJS-12 has strong measurement properties in terms of validity, internal consistency and sensitivity to change in TKA and THA patients. Low ceiling effects and good relative validity allow the monitoring of longer term outcomes, particularly in well-performing groups after total joint arthroplasty.

Cite this article: Bone Joint J 2017;99-B:218–24.

Total joint arthroplasty can achieve high levels of pain relief, function and satisfaction for patients with end-stage arthritis. In the United Kingdom, almost 100 000 hip and knee arthroplasty procedures are undertaken annually. The success of arthroplasty is assessed by patient-reported outcome measures (PROMs). Traditionally, joint-specific PROMs have evaluated domains such as pain and function. Patient demographics and expectations of outcome are important factors that influence the selection of PROMs.

Perhaps the ultimate aim of joint replacement surgery is for the patient to ‘forget’ that they have had surgery, or for the outcome to be so good that they become unaware of the (previously problematic) joint during daily life and activities. The 12 question Forgotten Joint Score (FJS-12) was introduced in 2012 with the aim of assessing the patients’ ‘joint awareness’. The authors of the score suggest this is more representative of higher-level function after surgery, as to be able to forget about the joint requires the absence of pain and the ability to perform all desired functional tasks without limitation. It is clearly important that an outcome score is psychometrically validated for use and known to accurately and reliably assess the factors that it pertains to. Other groups have used various language translations of the FJS-12 to evaluate outcomes in series of arthroplasty patients. A few authors have made limited attempts at psychometric validation, in particular the reliability of the score to achieve the same result has been evaluated and re-test reliability has been shown to be high, with intraclass correlation...
coefficients above 0.9. In general, previous studies have only assessed post-operative time points. Only one study, a French translation, has reported the pre-operative data that is essential to infer change from pre-operative score parameters and quantify operative success. To date, detailed psychometric validation of the score in a native English speaking population is lacking.

The aim of this study was to perform a comprehensive psychometric evaluation and investigate the measurement characteristics (dimensionality, validity and responsiveness) of the English language version of the FJS-12 in total hip and knee arthroplasty patients (THA and TKA). We also sought to validate longitudinal data in contrast to the routinely used Oxford Scores in a United Kingdom population.

**Patients and Methods**

**Participants.** Between January 2014 and August 2014, we prospectively assessed 1108 patients who were to undergo unilateral THA or TKA at a single NHS orthopaedic teaching hospital. The study centre is the only hospital to receive adult referrals in a predominantly urban regional population of approximately 850,000 people. Ethical approval was obtained for this project from the institutional review board (11/AL/0079); and patients were recruited with informed consent. All patient undergoing lower limb arthroplasty were eligible for the study. For analysis we required all three assessment time points, thus patients not providing complete follow-up assessment data were excluded. Individual patients could only be entered into the analysis once to prevent bias, and the second operation in cases of staged bilateral procedures was excluded. There were no other exclusion criteria.

A total of 546 patients underwent TKA and 562 patients underwent THA between January 2014 and August 2014. Prospective data was available for 467 TKA patients and 466 THA patients. A total of 22 TKA patients and 25 THA patients were excluded from the analysis as they underwent bilateral surgery within the study period. As such 445 TKA patients and 441 THA patients were eligible for the study. Of these, 214 TKA patients and 236 THA patients were excluded for not providing questionnaire data at all three assessment time points (pre-operatively, 6-months and 12-months). This resulted in 231 TKA and 205 THA patients eligible for the study.

Demographic data is presented in Table I. The mean age of THA cohort was 69.9 years with a 45.5% male and 55.5% female gender split. The mean age of the TKA cohort was 67.6 years with a 41.5% male and 58.5% female gender split.

Assessments took place at a pre-operative assessment clinic and then by postal questionnaire at six and
Assess the patient’s pain and function. Each item is answered on a five-point Likert response format. Item scores are summed and linearly transformed to a 0 to 100 scale, a high score (48) good joint function.

Joint awareness during the activities of daily living (for example, climbing stairs, walking for more than 15 minutes, in bed at night etc). It consists of 12 questions with a five-point Likert response format. Item scores are summed and linearly transformed to a 0 to 100 scale, a high value reflecting the ability of the patient to forget about the affected/replaced joint during the activities of daily living.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.

The SF-12 Health survey is a 12-item quality of life questionnaire often used in large population health surveys. Each item is rated using a five-point Likert response format. The transformed scores for each health domain reflect a score (PCS) and mental component summary score (MCS). Normative-based scoring is used for each component: the physical component summary (PCS) and generates a summed total score ranging from 0 to 48, used as a reference measure at all three time points as they have been proven to be responsive in the measurement of change.
Convergent and discriminant validity. To assess convergent validity we correlated the FJS-12 with the Oxford Scores and the SF-12 PCS. In TKA patients we found high correlations for the OKS ($r = 0.85$) and the SF-12 PCS ($r = 0.70$). Correlations in THA patients were slightly lower; $r = 0.79$ for the OHS and $r = 0.67$ for the SF-12 PCS. Discriminant validity was better in the TKA group (correlation with SF-12 MCS $r = 0.23$) than in the THA group (correlation with SF-12 MCS $r = 0.36$).

Sensitivity to change over time. To measure the performance of a score over time we analysed the data for all time-intervals following surgery (Figs 2 and 3). Results are detailed in Table II.

Pre-operative to six months post-operatively. Results for the FJS-12 showed a large ES for change from pre-operative to six-month follow-up in TKA patients (2.65) and in THA patients (2.27). ES for the OKS and OHS were 1.95 for both TKA patients and THA patients. RV for the FJS-12 was 2.27 in TKA patients and 2.01 in THA patients.

Six months to 12 months post-operatively. ES for the interval between the six month and one-year time questionnaires were smaller for FJS-12 than for the first follow-up (0.12 for TKA patients, 0.21 for THA patients). RV scores decreased to 0.31 in TKA patients and to 0.04 in THA patients. Results for the Oxford Scores showed an ES of 0.06 in TKA patients and -0.03 in THA patients.

Floor and ceiling effects. Pre-operatively, the FJS-12 showed pronounced floor effects, with 15.6% of the TKA patients and 22.4% of the THA patients obtaining the minimum score of 0 points. At one-year follow-up, 3.9% of TKA patients and 8.8% of THA patients achieved the maximum score of 100 points. In contrast, the Oxford Scores did not show floor effects. The ceiling effects for the

Table II. Means, standard deviations (sd) and responsiveness for the Forgotten Joint Score (FJS-12), Oxford Knee and Hip Scores (OKS and OHS) and the Mental and Physical Component Scores (MCS, PCS)

<table>
<thead>
<tr>
<th></th>
<th>Pre-operative</th>
<th>6 mths</th>
<th>1 yr</th>
<th>Pre-operative to 6 mths</th>
<th>6 mths to 1 yr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>sd</td>
<td>n</td>
<td>Mean</td>
<td>sd</td>
</tr>
<tr>
<td>Knee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FJS-12</td>
<td>11.5</td>
<td>11.6</td>
<td>231</td>
<td>42.1</td>
<td>30.0</td>
</tr>
<tr>
<td>OKS</td>
<td>20.7</td>
<td>7.4</td>
<td>231</td>
<td>35.0</td>
<td>8.8</td>
</tr>
<tr>
<td>MCS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PCS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hip</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FJS-12</td>
<td>12.3</td>
<td>15.9</td>
<td>205</td>
<td>48.4</td>
<td>28.7</td>
</tr>
<tr>
<td>OHS</td>
<td>21.0</td>
<td>9.0</td>
<td>205</td>
<td>38.6</td>
<td>8.6</td>
</tr>
<tr>
<td>MCS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PCS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* only administered at 12 months

ES, effect size; RV, relative validity
FJS-12 at one year post-operatively were approximately half of those seen in the Oxford Scores at the same time point (THA 18.5% FJS-12 versus 39.5% OHS, TKA 12.6% FJS-12 versus 25.5% OKS) (Table III).

**Missing responses.** Analysis of missing responses at a questionnaire item level highlighted generally high levels of completion. Item completion was 98% for the FJS-12 in the TKA cohort with the exception of question 12 (ability to forget about the joint when taking part in sports activities) which had missing responses of 12.1%. The same was found in the THA cohort, with more than 97% item completion for each question except item 12 concerning sporting activity, which was omitted in 19.8% of cases. The Oxford Scores did not suffer from missing responses, with more than 98% item completion across the 12 questions.

**Discussion**

‘Joint awareness’ is a distinct concept. Qualitative interviews with patients after knee arthroplasty have shown that joint awareness is triggered at a cortical level by pain, functional impairment, and associated sensations (such as numbness, lack of sensation, ‘strange feelings’ with touch, or weather-related sensations).27 Hudak et al28 define this as ‘disrupted embodiment’, where the joint has become separated from the body. From a theoretical standpoint, a high level of post-operative function may be related to the patient’s ability to ‘reunite’ the joint with the rest of the body and to become ‘unaware’ of it in everyday life, akin to the ‘normal’ condition of a ‘healthy’ joint.

With a variety of potential outcome assessment tools available, it is essential that detailed psychometric evaluation is available to inform choice of PROM instruments for the assessment of specific study endpoints. The Consensus-based Standards for the selection of health Measurement Instruments criteria suggest that PROM instruments should meet three main quality domains: reliability, validity and responsiveness.29,30

In this native English speaking population, the FJS-12 shows strong measurement properties in terms of dimensionality (reliability), convergent/discriminant validity and sensitivity to change over time (responsibility).

The Oxford Scores are regularly used tools that have been central to the development of arthroplasty outcomes. They are valid and reliable, although they may suffer from ceiling effects31,32 and may therefore lack the sensitivity to capture relevant changes in patient function over longitudinal time periods. As such there has been much interest in developing more sensitive tools with which to assess the outcome of arthroplasty.12,33,34

A major difference between the FJS-12 and the Oxford Scores is the population in which they were designed. The Oxford Scores were developed in pre-operative populations18,19.
and reflect the symptom state at that time point. In contrast, the FJS-12 was designed in a post-operative population and the construct reflects the symptom state after joint replacement surgery. This may explain the respective floor and ceiling effects of the scores. We have shown that the Oxford Scores are more responsive in reflecting the initial change between the pre-operative presentation and six months post-operatively, but less so for the change between six- and 12-month follow-up. As such, the FJS-12 had higher effect sizes than the Oxford Scores covering this time period, but lower relative validity. This is directly related to the floor effect of the FJS-12 pre-operatively which results in a low SD of the score at that time point. As the SD is the denominator of Cohen’s d, the effect sizes are inflated. By contrast, for the later follow-up period (six to 12 months), the FJS-12 outperformed the Oxford Scores in both THA and TKA cohorts. A recent Danish study also found that at later follow-up time points, the OKS has a higher ceiling effect than the FJS-12.14

To date, this is the most comprehensive evaluation of the psychometric properties of the FJS-12. Others have evaluated selected measurement characteristics in various cohorts and languages. Matsumoto et al7 found good convergent validity in a Japanese cohort of hip arthroplasty patients between the FJS-12 and the Western Ontario McMaster Universities Osteoarthritis index (WOMAC) (r = 0.53) and a culture-specific hip outcome instrument, the Japanese Orthopaedic Association Hip Disease Evaluation Questionnaire (r = 0.63). Thompson et al8 found a very high test-retest reliability for the FJS-12 (r = 0.97) in an Australian knee arthroplasty cohort. In this study, convergent validity was high with the WOMAC total score (r = 0.70) and Knee injury and Osteoarthritis Outcome Score (KOOS) and Quality of Life, Pain and activities of daily living scores (r = 0.65). They further reported less ceiling effect in the FJS-12 compared with the WOMAC and KOOS questionnaires. Similarly, Giesinger et al11 indicated superior responsiveness to change for the FJS-12 between one and two year follow-ups compared with the WOMAC and the Knee Society Score, using data from a Swiss clinical trial in knee arthroplasty patients. These three studies broadly support the results of this evaluation, where the FJS-12 showed strong convergent validity with the OKS and OHS and the physical component score of the SF-12. Reassuringly, the association with the mental health component of the SF-12 was low, highlighting appropriate divergent validity.

We report a lesser improvement in the FJS-12 from the pre-operative to one-year post-operative scores compared with Thienpont et al,6 who used a French translation of the score in a Belgian population. In their study of 75 THA and 75 TKA patients, there was a more pronounced ceiling effect post-operatively but less of a floor effect pre-operatively. This may highlight a variation in cultural response in international samples, or may also be due to higher rates of missing responses for individual FJS-12 items in that study. Reassuringly the baseline and post-operative scores in our sample, as measured by the Oxford Scores, are comparable with data from the United Kingdom National Joint Registry,35 suggesting our cohort data to be reflective of the population of the United Kingdom.

In our study, missing responses for individual FJS-12 items were below 3% for 11 out of 12 items in both hip and knee samples. A single question, that of joint awareness during sports, was less well reported. While this suggests that a modest proportion of United Kingdom arthroplasty patients do not consider a question on sporting participation to be applicable, over 80% did respond, perhaps reflecting the evolving expectation of higher level function of patients after joint arthroplasty.

The strengths of this study are its prospective design, a well-defined cohort of patients and the incorporation of pre-operative data. A potential limitation is that this study has a limited follow-up period of 12 months, although this is the typical time period reported in clinical outcome evaluations and in arthroplasty joint registries. The range of comparator outcome metrics is somewhat restricted. A further limitation is the percentage of patients with missing follow-up assessment data at six months or 12 months that were not included in our analysis. While this does not affect the validity of the analyses presented here, it may limit this data in terms of absolute clinical outcome scores.

The current study and previous work suggest that the FJS-12 has the benefits of improved responsiveness and reduced ceiling effects. The low ceiling effects of the FJS-12 may offer advantages when evaluating high performing groups or in powering clinical trials. We suggest that joint awareness could be seen as a relevant outcome domain in its own right, and complementary to questionnaires that assess pain and physical function. It may be that a battery of scores is required to evaluate the full impact of joint surgery from the point of view of the patient. As such, we suggest future studies evaluate qualitative aspects of the patient’s conceptualisation of ‘joint awareness’ and investigate the relative importance of joint awareness compared with joint-specific outcome domains (such as pain and function). Such questions could also be investigated quantitatively using structural equation modelling techniques. It may also be interesting to validate the FJS-12 for other joint conditions to investigate the potential of the score in patient groups where a good outcome following a surgical or conservative intervention is likely to be achieved, for example the conservative management of sports injuries.

The FJS-12 demonstrates strong measurement properties in terms of validity, uni-dimensionality and sensitivity to change in both hip and knee arthroplasty patients. Low ceiling effects and good relative validity suggest that this score may be beneficial for monitoring outcomes over the long-term in patient cohorts expected to achieve high levels of function, or for powering clinical trials.

VOL. 99-B, No. 2, FEBRUARY 2017
Take home message:
- This paper validates a new outcome score that assesses the patient's awareness of their joint replacement in a United Kingdom population.
- Strong measurement properties suggest this to be an interesting avenue for further research into patient outcomes, particularly in differentiating outcomes in highly performing patient groups.

Author contributions:
- D. F. Hamilton: Study design, Data analysis, Manuscript preparation/editing.
- F. L. Loth: Data analysis, Manuscript preparation/editing.
- J. M. Giesinger: Study design, Data analysis, Manuscript preparation/editing.
- K. Giesinger: Study design, Manuscript editing.
- D. J. Macdonald: Study design, Data collection, Manuscript editing.
- J. T. Patton: Study design, Manuscript editing.
- A. H. R. W. Simpson: Study design, Manuscript editing.
- C. R. Howie: Study design, Manuscript editing.

The author or one or more of the authors have received or will receive benefits for personal or professional use from a commercial party related directly or indirectly to the subject of this article.

This is an open-access article distributed under the terms of the Creative Commons Attribution licence (CC-BY-NC), which permits unrestricted use, distribution, and reproduction in any medium, but not for commercial gain, provided the original author and source are credited.

This article was primary edited by E. Moulder and first proof edited by A. C. Ross.

References