

# Interpreter proxy versus healthcare interpreter for administration of patient surveys following arthroplasty: a pilot study.

Daniel Xue (✉ [daniel.xue@health.nsw.gov.au](mailto:daniel.xue@health.nsw.gov.au))

Sydney Local Health District <https://orcid.org/0000-0001-8473-4762>

**Timothy Churches**

South Western Sydney Local Health District

**Elizabeth Armstrong**

University of New South Wales

**Rajat Mittal**

South Western Sydney Local Health District

**Justine M. Naylor**

South Western Sydney Local Health District

**Ian A. Harris**

South Western Sydney Local Health District

---

## Research article

**Keywords:** Osteoarthritis, Arthroplasty, Clinical Quality Registry, Interpreter, Reliability,

**Posted Date:** October 21st, 2019

**DOI:** <https://doi.org/10.21203/rs.2.12241/v3>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published on November 14th, 2019. See the published version at <https://doi.org/10.1186/s12874-019-0854-1>.

# Abstract

**Background:** Clinical quality registries and other systems that conduct routine post-discharge surveillance of patient outcomes following surgery may have difficulty surveying patients who have limited proficiency in the language of the healthcare provider. Interpreter proxies (family and carers) are often used due to limited access to certified healthcare interpreters (due to cost or availability). The aim of this study was to assess the reliability of engaging interpreter proxies compared with certified healthcare interpreters for the administration of patient-reported health-related surveys for people with limited English proficiency (LEP). **Methods:** People with LEP and due for a routine 6-month telephone follow-up post knee or hip arthroplasty were invited to participate. Participants were randomly allocated to having their first interview with an interpreter proxy or a certified healthcare interpreter followed by the second (crossover) interview within 2 weeks (range: 4 to 12 days) after the first interview using the alternative method. Agreement between the two methods was assessed using quadratic weighted Cohen's kappa, intraclass correlation and concordance correlation coefficient where appropriate for EQ-5D health domains, total Oxford hip and knee scores, patient satisfaction, operation success, readmission, reoperation, and post-surgical complication responses. The mean of the differences between the same data items collected by each of the two methods was also calculated. **Results:** 85 participants (96%) completed the study. There was substantial to excellent inter-rater agreement (kappa=0.69-0.87 and ICCs above 0.74) for all but one measure. The mean differences between family proxy and healthcare interpreter scores for each participant were small, ranging from 0.01 (score range of 1-5) to 0.72 (score range of 0-100). **Conclusion:** These results suggest that using interpreter proxies is a reliable alternative to certified healthcare interpreters in conducting patient-reported health surveys, potentially making this process easier and cost effective for researchers and registries.

## Introduction

Hip and knee arthroplasties are common surgical procedures and patients are commonly surveyed post-operatively to ascertain their health outcomes after treatment. A major challenge to the collection of such data is inclusion of people with limited or no language proficiency in the language of the healthcare provider or the surveyor. For those with limited English proficiency (LEP), for example, their satisfaction with healthcare may be lower, and they may have a greater risk of serious medical events (1). In an Australian study, non-English speaking status was a predictor of lower postoperative International Knee Society scores and more severe self-reported pain following knee arthroplasty (2). These observations suggest it is imperative that LEP patients have their health outcomes included in surveillance programs if the outcomes are to reflect the entire population.

For many outcome measurement programs, patients are surveyed by telephone. The costs of professional interpreter services (3) and logistical difficulties of finding multi-lingual staff to administer questionnaires in the patients' native languages means that family or carers (who speak the requisite language) are often used as interpreters (termed as interpreter proxies). However, the reliability of using interpreter proxies in this setting has not been compared to the use of professional interpreters.

The aim of the study is to determine whether, in patients with LEP, the use of interpreter proxies provides adequately reliable survey results compared to using certified healthcare interpreters, which are considered the 'gold standard'.

## Methods

We used a randomised crossover study design to compare survey outcomes between two groups: surveys conducted using interpreter proxies (family members and carers), and those conducted using certified healthcare interpreters.

The study setting was within the Arthroplasty Clinical Outcomes Registry National (ACORN), a clinical quality registry that collects health data on patients undergoing elective hip or knee arthroplasty surgery in multiple hospitals. Post-operative data collection is conducted 6 months post-surgery by telephone and approximately 12% of participants have LEP.

The participants included in this study were ACORN patients who were due for their 6-month follow-up call between March and September 2015. Inclusion criteria were: having identified themselves as requiring an interpreter and/or having LEP preoperatively, cognitive capacity to understand the follow-up questions, fluency in a language for which a healthcare interpreter is available from the South West Sydney Interpreter Services, and also having access to a family or friend proxy who is able to interpret between English and their desired language.

English proficiency was ascertained in the pre-admissions clinic through a sensitive set of questions (4). Patients were asked *how well do you speak English?* Those who answered *very well* were deemed English proficient while those who answered *well* or *not well* were asked a second question: *In what language do you prefer your medical care?* Those nominating 'English' as their preferred language, were classified as English-proficient. Those nominating another language or who were unable to answer the first question were classified as having LEP and were included in the screening process for this study.

Patients with LEP who met the above criteria were sequentially ordered according to when their interviews were due and were contacted by telephone. If patients provided verbal consent with the assistance of the healthcare interpreter or interpreter proxy, they were randomly allocated to having their first interview via an interpreter proxy or via a certified healthcare interpreter. The randomisation was carried out according to a computer-generated sequence prepared before the commencement of the study and concealed in sequentially numbered envelopes containing allocation details. The envelopes were opened immediately after the patient provided consent.

The first interview was conducted within one month either side of the 6 month post-operative date as per routine follow up. We considered the condition to be stable at 6 months. This was followed by the second interview within 2 weeks of the first interview using the alternative method. The interview questions were asked in English by the research officer and questions and responses were translated to and from the appropriate language by the interpreter proxy or certified healthcare interpreter. Interviews with certified

interpreters were made with the assistance of the call centre manager at the Sydney South West Local Health District Language Services who connected the research officer, interpreter and patient in a 3-way conference call. Interviews with interpreter proxies were performed in a 2-way telephone call with the research officer on one end and the proxy translating the interview questions and responses to and from the patient on the other end.

The questions asked were the standard 6-month follow-up questions for all patients in the ACORN registry for determining patient-reported outcome measures (PROMs) as shown below:

1. Satisfaction: "How would you describe the results of your operation" (a 5-point Likert scale - 'excellent'-1, 'very good'-2, 'good'-3, 'fair'-4, 'poor'-5).
2. Success: "Overall, how are the problems with your knee/hip now compared to before your operation" (a 5-point Likert scale - 'much better'-1, 'a little better'-2, 'about the same'-3, 'a little worse'-4, and 'much worse'-5).
3. Complications: "Have you experienced any complications after the operation since being discharged from hospital"; a standard list of common complications was read out.
4. Readmission: "Were you admitted to hospital again since leaving hospital after the knee/hip replacement?" answered as yes or no.
5. Reoperation: "Have you had another operation on the same joint that was operated on?" answered as yes or no.
6. Patient-reported health status using the EuroQoL EQ-5D-5L and EQ-VAS questionnaires (5) (English version): The EQ-5D-5L rates the patient's mobility, personal care, usual activities, pain/discomfort and anxiety/depression levels in separate 5-point Likert scales, in which for each category a score of '1' represents the best outcome and a score of '5' the worst. The EQ-VAS rates the patient's overall health along a visual scale from zero to 100 where zero refers to the worst health and 100 the best health. The English version of the questionnaires were used and read out by the interpreter and interpreter proxies in the patient's desired language.
7. Joint-specific patient-reported pain and function was assessed using the Oxford Hip Score (OHS) and Oxford Knee Score (OKS) (English version). This is a 12-question survey using a Likert scale (0-4) which asks about the patient's perceived difficulty or pain with performing everyday movements and tasks. The summary score minimum is 0 and the maximum score of 48 denotes the best outcome (6, 7). English versions of the Oxford scores were used and read out by the interpreters and interpreter proxies in the patient's desired language.
8. Two extra questions were asked to determine the number of years the patient and interpreter proxy had spent living in Australia: "*in what year did you (and interpreter proxy) first arrive in Australia to live here for one year or more?*" These questions correspond with those asked in the 2011 Australian Census of Population and Housing (8).

Our convenience sample exceeded the minimum sample size of 50 patients required to detect an ICC of 0.50 with 90% power and 5% significance. Ordinal data were analysed using quadratic weighted Cohen's

kappa coefficients, intraclass coefficients (ICC) measuring absolute agreement and Lin's concordance correlation coefficient (CCC) for each outcome measure to assess the magnitude of agreement between interpreter proxy and healthcare interpreter. In addition, a Wilcoxon paired ranked sum test was performed on each measure to assess the statistical significance of the differences obtained between the two methods of interview administration. Nominal data measures were analysed with an unweighted Cohen's kappa coefficient. The EQ-VAS was treated as continuous data and was analysed using ICC and CCC.

The differences in scores from the two methods were also determined for the EQ-5D, EQ-VAS, and Oxford scores and visualised through Bland-Altman plots (9). From this a mean was calculated to evaluate for potential bias. The degree of bias is revealed by the mean differences and 95% Limits of Agreement plotted to indicate the range where 95% of the differences lie. All data analysis was performed using R open-source statistical software version 3.2 (10). Figures were generated using R Studio version 0.99.

The outcomes assessed were the levels of agreement between the two methods of language interpreting as determined by Cohen's *kappa* coefficients, Intraclass correlation coefficient (ICC) and Concordance correlation coefficient (CCC) statistics where appropriate. The Cohen's kappa coefficients were interpreted in accordance with guidelines put forward by Landis and Koch (11). Coefficients between 0.21 and 0.40 were considered to show fair agreement, scores between 0.41 and 0.60 moderate agreement, scores between 0.61 and 0.80 substantial agreement, and scores above 0.80 almost perfect agreement.

## Results

125 of the patients due for follow up calls between March 2015 and September 2015 were screened and as LEP and invited to participate. Out of these, 89 patients provided consent and were given a random allocation of call order ( $n=46$ , interpreter proxy first;  $n=43$ , health care interpreter first). The 36 exclusions were due to not having an interpreter proxy available for translation, speaking a language for which there was no healthcare interpreter available, or if the patient was not contactable for either interview method. Four patients out of the 89 (two from each arm) withdrew from the study after consent was obtained and completion of the first interview. 85 patients successfully received both methods of follow-up calls and were included in the data analysis. Figure 1 summarises the flow of patients throughout the study. Descriptive demographics were similar in both arms of the study and are displayed in Table 1 and the languages used are presented in Table 2.

**Table 1: Demographic profile of the study population.**

|                                       |        | Total<br>(n=85) | Int Proxy First (n=44) | HC Int First (n=41) |
|---------------------------------------|--------|-----------------|------------------------|---------------------|
| Sex, n (%)                            | Male   | 25 (29)         | 13 (15)                | 12 (14)             |
|                                       | Female | 60 (71)         | 31 (36)                | 29 (34)             |
| Age, mean (SD)                        |        | 72 (7.7)        | 72 (7.1)               | 71 (8.5)            |
| Patient years in Australia, mean (SD) |        | 30 (17)         | 31 (16)                | 28 (17)             |
| Proxy years in Australia, mean (SD)   |        | 31 (15)         | 32 (15)                | 30 (15)             |
| Surgery Type, n (%)                   | Knee   | 66 (78)         | 37 (44)                | 29 (34)             |
|                                       | Hip    | 19 (22)         | 7 (8)                  | 12 (14)             |

*Int Proxy: Refers to family or carers of patients acting as interpreters*

*HC Int: Certified healthcare interpreters*

*SD: standard deviation*

**Table 2 – Languages utilised in this study**

| Language   | N  | Mean age | Mean years in Australia |
|------------|----|----------|-------------------------|
| Arabic     | 16 | 68.5     | 27.2                    |
| Spanish    | 11 | 76.6     | 33.2                    |
| Chinese    | 11 | 71.8     | 17.8                    |
| Greek      | 8  | 77.6     | 51.1                    |
| Macedonian | 7  | 68.9     | 38.0                    |
| Italian    | 6  | 76.0     | 40.3                    |
| Serbian    | 4  | 77.5     | 36.2                    |
| Vietnamese | 4  | 68.5     | 27.5                    |
| Assyrian   | 3  | 67.7     | 4.0                     |
| Punjabi    | 3  | 69.7     | 8.3                     |
| Croatian   | 2  | 76.5     | 49.0                    |
| Farsi      | 2  | 59.0     | 10.0                    |
| Others*    | 8  | 68.4     | 26.6                    |

*Preferred languages for the included patients and the mean number of years the patient had spent living in Australia.*

*\*One each for Armenian, Croatian, Haka, Maltese, Portuguese, Tamil, Turkish, Urdu*

Agreement between methods of interview was at least substantial (agreement score >0.60) for all outcomes except for the anxiety/depression section of EQ-5D which scored 0.57 as shown in Table 3. The remainder of the kappa scores ranged from 0.66 to 1, ICCs ranged from 0.66 to 0.87 and CCCs from 0.66 to 0.87. The CCC Plots for each domain of the EQ-5D are depicted in table 2. Perfect agreement was seen in re-admission and re-operation responses and almost perfect agreement was seen in the Oxford score responses (*kappa and ICC=0.87, CCC=0.86*). Agreement for different sections of the EQ-5D varied considerably with scores from 0.57 anxiety/depression to 0.81 for mobility. The CCC Plots for each domain of the EQ-5D are depicted in Figure 2. Patients who underwent hip arthroplasty consistently demonstrated a higher level of agreement when compared with knee arthroplasty patients.

The CCC plots for the EQ VAS and Oxford scores (Figures 3 and 4 respectively) show the scale and location shifts for the CCC analyses are minimal with the scale shifts ranging from 0.89 to 1.16 and

location shifts from -0.05 to -0.02.

**Table 3** – Agreement scores of patient reported outcome measures

|                    | <i>kappa</i> | ICC (95% CI)     | CCC (95% CI)       | MD    | 95% Limits of Agreement | Wilcoxon <i>p</i> -value |
|--------------------|--------------|------------------|--------------------|-------|-------------------------|--------------------------|
| EQ-5D              |              |                  |                    |       |                         |                          |
| Mobility           | 0.81         | 0.82 (0.73-0.88) | 0.81 (0.73 - 0.87) | -0.02 | -0.93 to 0.88           | 0.66                     |
| Personal care      | 0.66         | 0.66 (0.52-0.77) | 0.66 (0.52 - 0.76) | 0.02  | -1.11 to 1.15           | 0.84                     |
| Usual Activities   | 0.68         | 0.68 (0.54-0.78) | 0.68 (0.54-0.77)   | 0.11  | -0.99 to 1.2            | 0.09                     |
| Pain/Discomfort    | 0.69         | 0.69 (0.56-0.79) | 0.69 (0.56-0.79)   | -0.02 | -1.07 to 1.02           | 0.70                     |
| Anxiety/Depression | 0.57         | 0.57 (0.41-0.7)  | 0.57 (0.41-0.69)   | 0.06  | -1.05 to 1.16           | 0.38                     |
| EQ-VAS             | -            | 0.78 (0.68-0.85) | 0.78 (0.68 - 0.85) | 0.72  | -22.7 to 24.1           | 0.79                     |
| Oxford scores      | 0.87         | 0.87 (0.8-0.91)  | 0.86 (0.8-0.91)    | 0.13  | -6.46 to 6.72           | 0.79                     |
| Satisfaction       | 0.75         | -                | -                  | 0.01  | -                       | 0.86                     |
| Success            | 0.70         | -                | -                  | -0.02 | -                       | 0.71                     |
| Re-admission       | 1.00         | -                | -                  | -     | -                       | -                        |
| Re-operation       | 1.00         | -                | -                  | -     | -                       | -                        |
| Complications      | 0.69         | -                | -                  | -     | -                       | -                        |



*Proxy: Interpreter proxy (family or carer)*

*HC Int: Certified healthcare interpreter*

*Kappa: Quadratic Weighted Cohen's Kappa*

*ICC: Intraclass correlation*

*CCC: Concordance correlation coefficient*

*MD: Mean difference between the two measurements (Proxy score minus HC Int score)*

*95% Limits of Agreement: MD ± 1.96 SD of the MD*

*Score Range for EQ5D: 1-5 for each domain, EQ- VAS: 0-100, Oxford Score: 0-48, Satisfaction: 1-5,*

*Success: 1-5.*

---

The Bland-Altman plots (Figures 5 and 6) show that relative to the total score range for each measure, the limits of agreement are narrower for the Oxford score compared with EQ-VAS. Overall the mean difference is very small (up to 1.1% for the Oxford) indicating negligible bias when all subjects are considered.

The Wilcoxon rank sum tests indicated that the differences between mean scores from either method of interview were not statistically significant. When results were stratified according the interview order, the differences between mean scores remained insignificant except for the EQ-VAS mean difference where if the first call was made with a healthcare interpreter this resulted in a slightly higher score (Table 4).

**Table 4:** Mean differences stratified by interview order

|                    | Interview Order      | MD    | 95% Limits of Agreement | Wilcoxon <i>p</i> -value |
|--------------------|----------------------|-------|-------------------------|--------------------------|
| Mobility           | Proxy first          | 0.02  | -0.87 to 0.91           | 0.82                     |
|                    | HC Interpreter first | -0.07 | -0.99 to 0.85           | 0.35                     |
| Personal care      | Proxy first          | -0.02 | -1.25 to 1.21           | 0.61                     |
|                    | HC Interpreter first | 0.07  | -0.94 to 1.09           | 0.41                     |
| Usual Activities   | Proxy first          | 0.11  | -1.17 to 1.40           | 0.31                     |
|                    | HC Interpreter first | 0.10  | -0.76 to 0.95           | 0.20                     |
| Pain/Discomfort    | Proxy first          | -0.14 | -1.22 to 0.95           | 0.12                     |
|                    | HC Interpreter first | 0.10  | -0.86 to 1.06           | 0.24                     |
| Anxiety/Depression | Proxy first          | 0.09  | -1.16 to 1.35           | 0.43                     |
|                    | HC Interpreter first | 0.02  | -0.90 to 0.95           | 0.82                     |
| EQ-VAS             | Proxy first          | 3.73  | -25.20 to 32.60         | 0.11                     |
|                    | HC Interpreter first | -2.51 | -15.70 to 10.60         | 0.04                     |
| Oxford scores      | Proxy first          | 0.64  | -7.57 to 8.84           | 0.20                     |
|                    | HC Interpreter first | -0.42 | -4.48 to 3.65           | 0.15                     |

*Proxy: Interpreter proxy (family or carer)*

*HC Int: Certified healthcare interpreter*

*Kappa: Quadratic Weighted Cohen's Kappa*

*ICC: Intraclass correlation*

*CCC: Concordance correlation coefficient*

*MD: Mean difference between the two measurements (Proxy score minus HC Int score)*

*95% Limits of Agreement: MD ± 1.96 SD of the MD*

*Score Range for EQ5D: 1-5 for each domain, EQ- VAS: 0-100, Oxford Score: 0-48,*

*Satisfaction: 1-5, Success: 1-5.*

---

## Discussion

The high level of agreement overall between interpreter proxies and healthcare interpreters found in this study suggests that utilising carers or family members to interpret is adequately reliable for capture of patient-reported outcomes after arthroplasty.

Healthcare interpreters have been noted to deliver a higher standard of interpreting quality with fewer translation errors when compared with non-certified *ad hoc* interpreters (12). As such, the healthcare interpreters should be adept at clarifying any confusion over the meaning of certain words in the questionnaire and also clearly relay how the patient's recovery profile is represented by the responses given.

The perfect agreement ( $kappa=1$ ) between the responses regarding readmission and reoperation was expected assuming that the question was accurately asked by both types of interpreters as these are significant events for the patient. Agreement for determining complication was lower ( $kappa=0.69$ ), a discrepancy which may have been due to uncertainty over the exact nature of a complication, which was defined as requiring active management but not readmission or reoperation.

The results showed that patients who had a hip arthroplasty recorded higher levels of agreement for all outcomes apart from complication, despite the only differences in questioning being the three joint-specific items in the Oxford questionnaires. This may have been due to the fact that patients who undergo knee operations experience greater day-to-day variability in their pain and function levels (possibly even dependent on weather conditions), which would account for the lower agreement levels in the EQ-5D which assesses daily situation compared with the Oxford scores which assess the most recent 4 week period.

Similar 95% limits of agreement have also been seen in test-retest studies of reliability indicating that the differences seen between methods in this study may be explained by the normal week-to-week variation of responses to the survey questions, which may also incorporate true health changes between surveys (13). While patient recovery following arthroplasty largely plateaus after 6 months, the one month window either side of the 6 month date which we allowed for conducting interviews may have confounded the test-retest reliability.

However, once the entire sample population is considered, the overall bias or the mean difference on the Bland-Altman plots is consistently very small. With results stratified to interview order, there was no statistically significant difference in the mean scores except for marginal differences in the EQ-VAS score. This indicates that patients were not continuing to improve in the time between interviews. This suggests that the methods are similar when comparing groups of patients but that differences are seen at the level of the individual patient. However, as stated above, we consider this to be reflective of the reliability of the survey questions rather than due to the method of data collection.

A limitation of the study was the inability to assess the effect that individual languages may have on the level of agreement. This was due to an insufficient sample size and the large number of languages, which did not allow analysis of individual languages. While languages could potentially be grouped into

geographical regions of origin, these classifications do not necessarily reflect cultural diversity and the effects this may have on patient reported outcomes. The study was limited to patients who had undergone arthroplasty surgery and these findings may not be generalisable to other settings, particularly where socially sensitive topics may be discussed which may limit the accuracy of proxy interpreters.

Another limitation is that we used English versions of the validated surveys and not those specific to the language of the patient being interviewed. Due to this, variation in the linguistic skills of proxy interpreters may have affected the accuracy of results. This approach, necessitated by a lack of availability of all required languages for each survey at the time of the study, reflected the pragmatic approach used by this registry that removes the need to restrict participation based on survey availability.

## **Conclusion**

The benefits of professional interpreters to the patient in a clinical setting are well recognised, with the literature agreeing that patient satisfaction and quality of care is greatest when hospital-trained interpreters and telephone interpreters are utilised (14-16). However, the inconvenience and cost of using healthcare interpreters is a barrier to participation in data collection.

These results suggest that using interpreter proxies is a reliable, efficient and likely cost-effective alternative to using healthcare interpreters when conducting telephone surveys of patient-reported outcomes after health interventions. Despite differences seen at the individual level, when the entire cohort is considered, there is an insignificant difference between the two methods of interview.

## **Abbreviations**

ACORN: Arthroplasty Clinical Outcomes Registry National

CCC: Concordance correlation coefficient

ICC: Intraclass correlation coefficient

Kappa: Quadratic Weighted Cohen's Kappa

LEP: Limited English proficiency

OHS: Oxford Hip Score

OKS: Oxford Knee Score

PROMs: Patient-reported outcome measures

## **Declarations**

**Ethics approval and consent to participate**

The study was approved by Hunter New England Human Research Ethics Committee and all procedures performed were in accordance with the ethical standards of the 1964 Helsinki declaration.

Informed consent was obtained verbally from all individuals included in the study with the assistance of certified interpreters or carers and family members. A verbal consent form approved by Hunter New England Human Research Ethics Committee was read out and completed for each participant enrolled in the study.

### **Consent for publication**

Not Applicable

### **Availability of data and materials**

The datasets analysed during the current study are available from the corresponding author on reasonable request.

### **Competing interests**

The authors declare that they have no competing interests

### **Funding**

Funding for the engagement of certified healthcare interpreters of Sydney South West Local Health District Language Services was provided by internal institutional funds of the Whitlam Orthopaedic Research Centre. No additional funding was provided for the design of the study, collection, analysis, and interpretation of data and in writing the manuscript for this study.

### **Authors' contributions**

EA, JN and IAH conceived and designed the study. EA submitted the study protocol for ethics approval. DX, TC, EA and RM contributed to data collection. DX and TC performed the data

### **Acknowledgements**

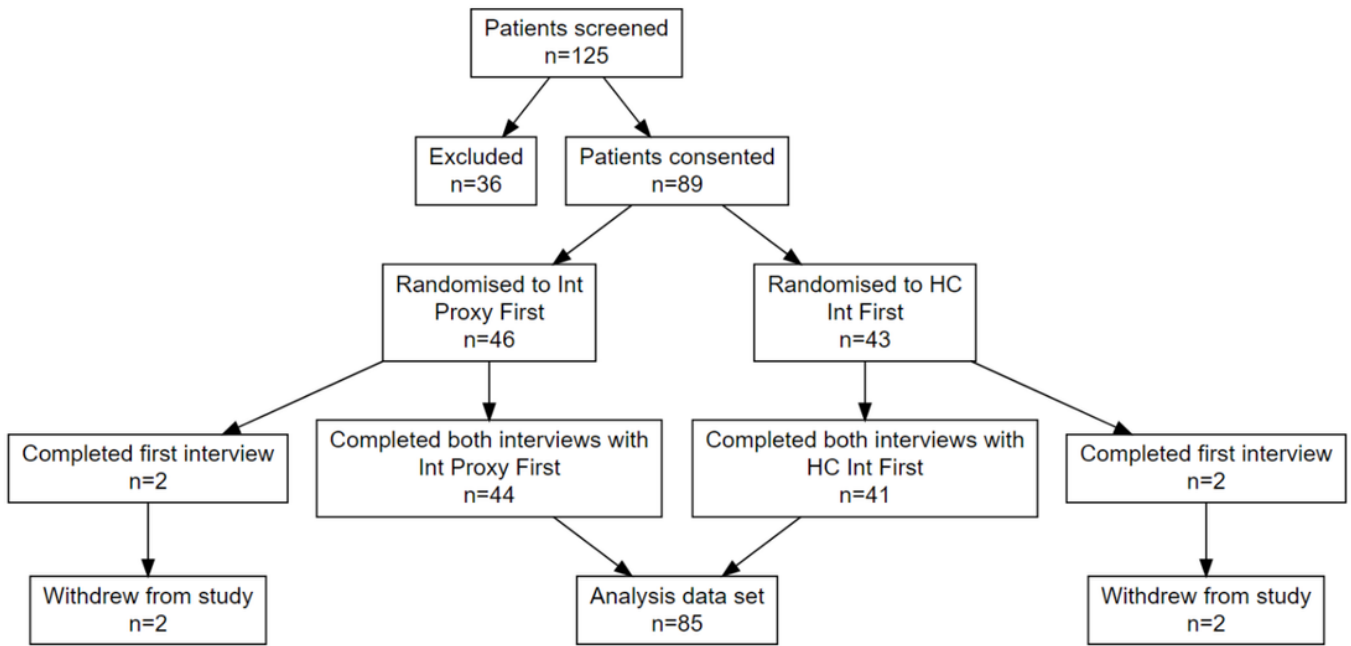
The authors wish to thank Katina Varelis, Nitin Sharma, Zareena Azimullah and the interpreter service of the Sydney South West Local Health District Language Services for organizing interpreter calls throughout the project.

## **References**

1. Cohen AL, Rivara F, Marcuse EK, McPhillips H, Davis R. Are language barriers associated with serious medical events in hospitalized pediatric patients? *Pediatrics*. 2005;116(3):575-9.

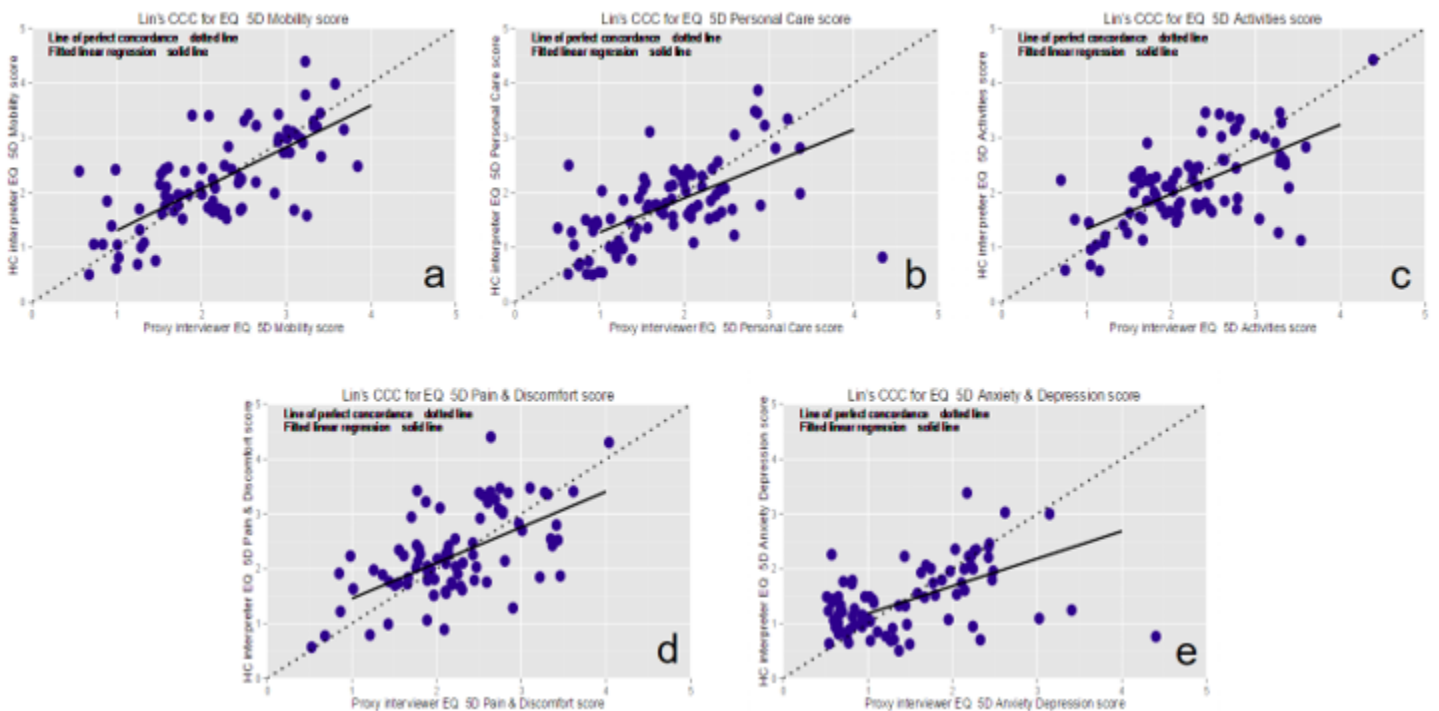
2. Dowsey MM, Broadhead ML, Stoney JD, Choong PF. Outcomes of total knee arthroplasty in English-versus non-English-speaking patients. *Journal of Orthopaedic Surgery*. 2009;17(3):305-9.
3. Jacobs EA, Shepard DS, Suaya JA, Stone E-L. Overcoming language barriers in health care: costs and benefits of interpreter services. *American journal of public health*. 2004;94(5):866-9.
4. Karliner LS, Napoles-Springer AM, Schillinger D, Bibbins-Domingo K, Pérez-Stable EJ. Identification of limited English proficient patients in clinical care. *Journal of general internal medicine*. 2008;23(10):1555-60.
5. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of life research*. 2011;20(10):1727-36.
6. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *Journal of Bone & Joint Surgery, British Volume*. 1996;78(2):185-90.
7. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. *Journal of Bone & Joint Surgery, British Volume*. 1998;80(1):63-9.
8. Australian Bureau of Statistics. 2011 Census Household Form – text version 2013 [Available from: <http://www.abs.gov.au/websitedbs/censushome.nsf/home/2011hhftranscript%5C>].
9. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical methods in medical research*. 1999;8(2):135-60.
10. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria 2015.
11. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33(1):pp. 159-74.
12. Karliner LS, Jacobs EA, Chen AH, Mutha S. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. *Health services research*. 2007;42(2):727-54.
13. Naylor JM, Kamalaseena G, Hayen A, Harris IA, Adie S. Can the Oxford Scores be used to monitor symptomatic progression of patients awaiting knee or hip arthroplasty? *The Journal of arthroplasty*. 2013;28(9):1454-8.
14. Garcia EA, Roy LC, Okada PJ, Perkins SD, Wiebe RA. A comparison of the influence of hospital-trained, ad hoc, and telephone interpreters on perceived satisfaction of limited English-proficient parents presenting to a pediatric emergency department. *Pediatric Emergency Care*. 2004;20(6):373-8.
15. Kuo D, Fagan MJ. Satisfaction with methods of Spanish interpretation in an ambulatory care clinic. *Journal of General Internal Medicine*. 1999;14(9):547-50.
16. Lee LJ, Batal HA, Maselli JH, Kutner JS. Effect of Spanish Interpretation Method on Patient Satisfaction in an Urban Walk-in Clinic. *Journal of General Internal Medicine*. 2002;17(8):641-6.

## Figures



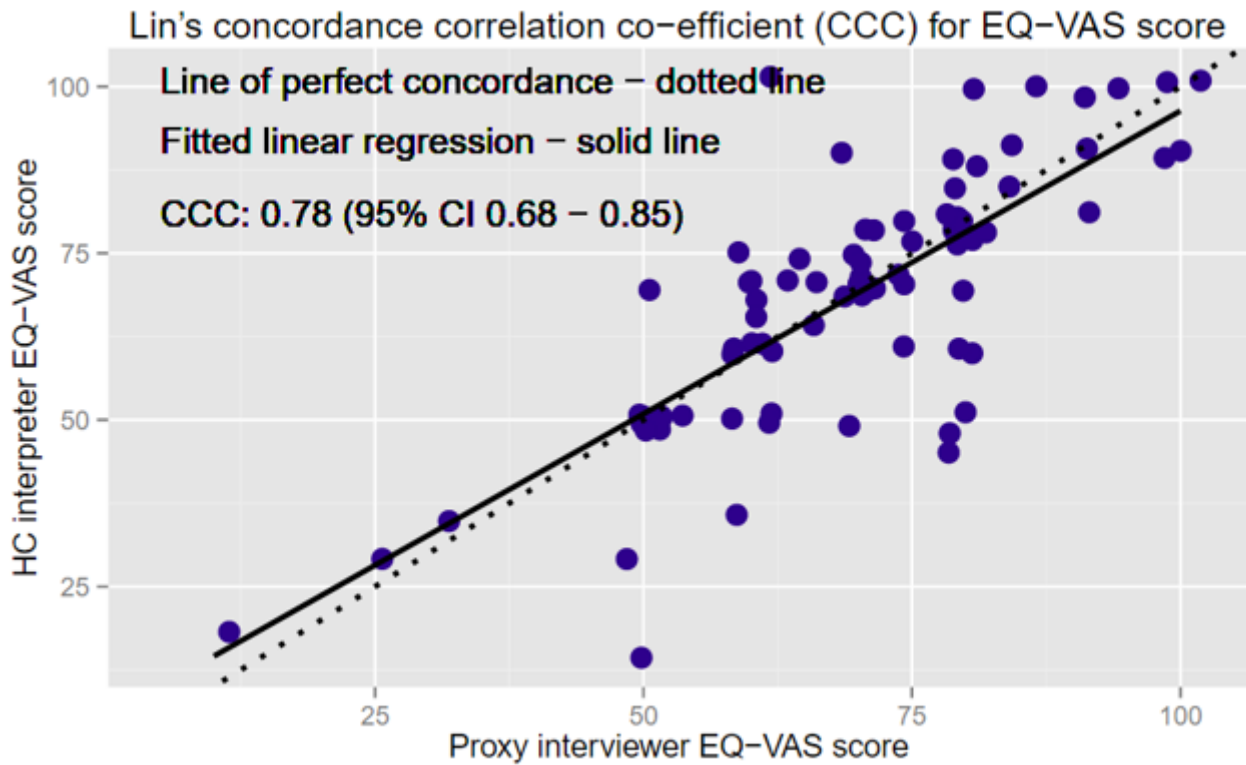
**Figure 1**

Patient flow diagram. 125 patients were identified as LEP between March 2015 and September 2015. Of these 89 were consented and randomly allocated to a call order. 46 were allocated to having the first call performed with an interpreter proxy while 43 were allocated to a healthcare interpreter first. 4 patients withdrew from the study after randomisation resulting in 85 patients being included in the final data analysis. Int Proxy: Refers to family or carers of patients acting as interpreters HC Int: Certified healthcare interpreters



**Figure 2**

CCC plots of the separate components of EQ-5D-5L with the score from proxy interviewer plotted against healthcare interpreter: (a) mobility (CCC=0.81), (b) personal care CCC=0.66), (c) usual activities CCC=0.68), (d) pain/discomfort CCC=0.69), and (e) anxiety/depression CCC=0.57). A fitted linear regression (solid) is compared with a 45° line (dotted) through the origin for each plot.



**Figure 3**

CCC plot of the EQ-VAS score (CCC=0.78).

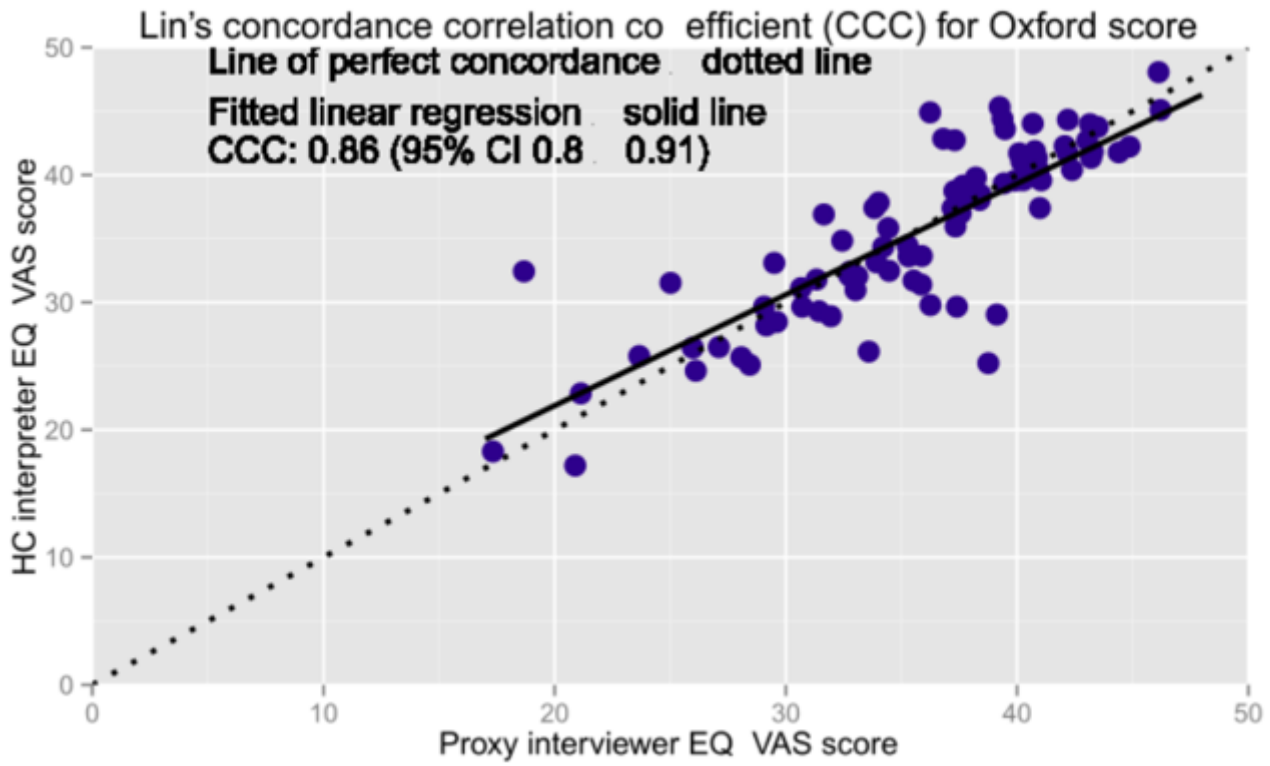
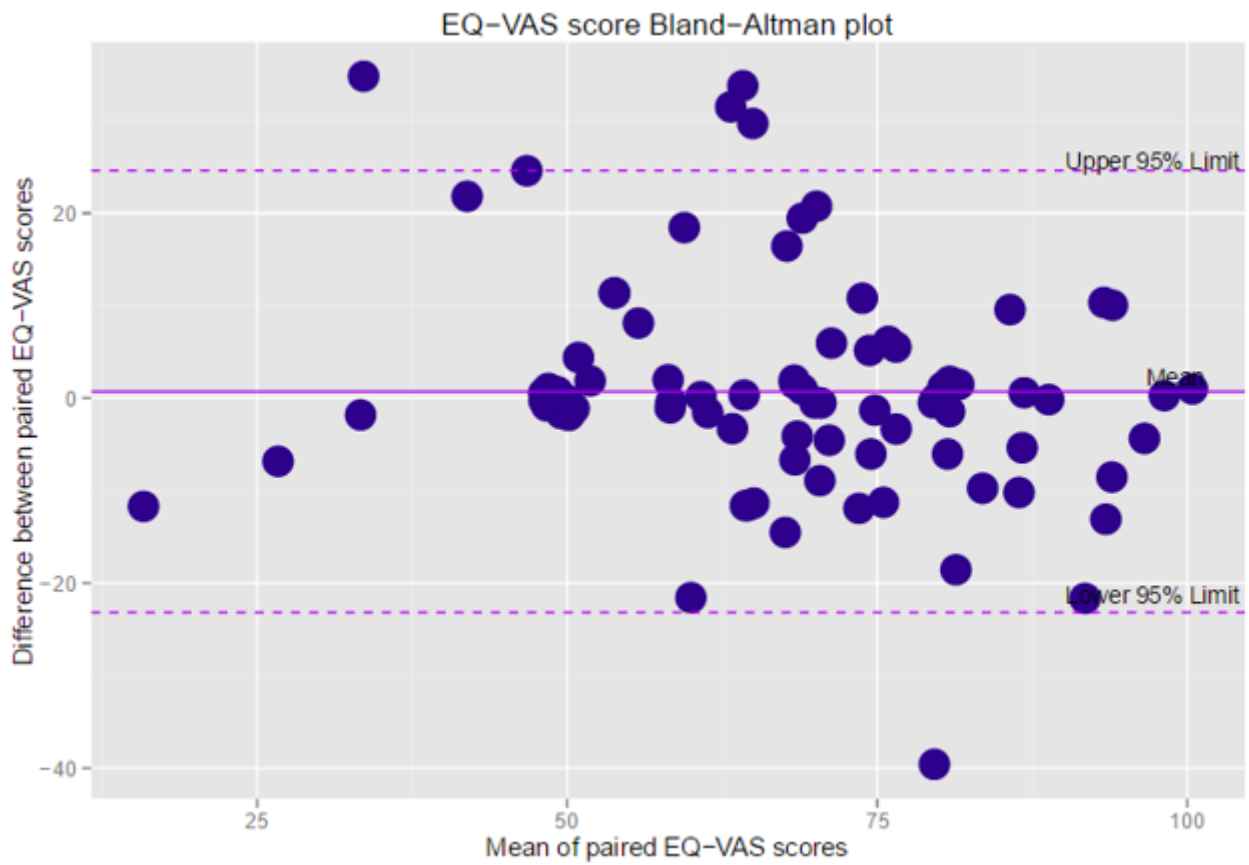


Figure 4

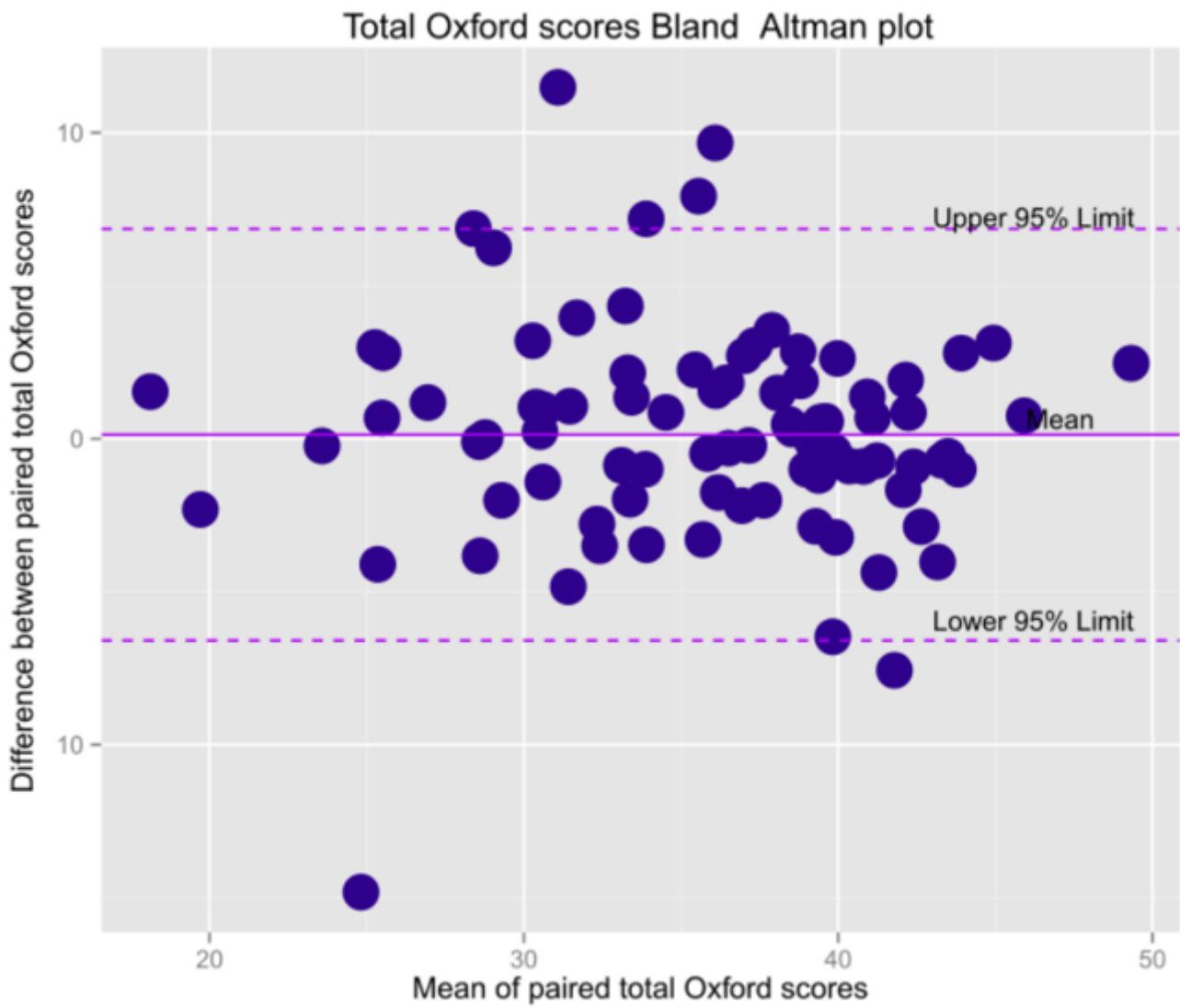
CCC plot of the Oxford hip/knee score (CCC=0.86). Data from both the hip and knee questionnaires have been analysed together.





**Figure 5**

Bland-Altman plot of the EQ-VAS score. The mean of both measurements is plotted against the difference between measurements (proxy interviewer score minus healthcare interpreter score). Mean difference = 0.72 with 95% limits of agreement -22.7 to 24.1.



**Figure 6**

Bland-Altman plot of the Oxford hip/knee score. Mean difference = 0.13 with 95% limits of agreement -6.46 to 6.72.