The test-retest reliability of pain outcome measures in people with phantom limb pain.

Authors: Andrew N. Graham[a] , Cormac G. Ryan[a], Alasdair MacSween[a], Greg Atkinson[b], Sally Smith[c], Denis J. Martin[a,d]

Affiliation: [a] Centre for Rehabilitation, School of Health and Life Sciences, Teesside University, United Kingdom, [b]School of Sport and Exercise Sciences, Liverpool John Moores University, UK [c]South Tees Hospitals NHS Foundation Trust [d] NIHR Applied Research Collaborative, North-East and North Cumbria, United Kingdom

Conflicts of Interest and Source of Funding: DM and CR are named inventors on a medical device for the management of phantom limb pain. They may benefit financially if the device is commercialised. This work was undertaken as part of the PhD of AG, which was primarily funded by Teesside University but also received additional funding from Medi-Direct Ltd. who were involved in the development of the aforementioned medical device for the management of phantom limb pain. None of the other authors have any conflicts of interest to declare.

Correspondence: Andrew Graham, School of Health & Life Sciences, Centuria Building, Teesside University, TS13BX, UK. andrew.graham@tees.ac.uk

Abstract

Objectives: To quantify the test-retest reliability of three patient-reported outcome measures of pain for people living with phantom limb pain (PLP) and assess impact of test-retest errors on future research and clinical decisions.

Methods: Thirty-nine participants (30 males), mean (SD) age: 55 (16), mean (SD) years post amputation: 6.8 (8.3), reported their PLP levels on a Visual Analogue Scale (VAS) for pain intensity, the revised Short Form McGill Pain Questionnaire (SF-MPQ-2), and a pain diary, on two occasions 7-14 days apart. Mean systematic change, within-subjects SD, limits of agreement (LOA), coefficient of variation and the intraclass correlation coefficient (ICC) were quantified alongside their respective 95% confidence intervals (95%CI).

Results: Systematic learning effects (mean changes) were not clinically relevant across the VAS, SF-MPQ-2 and pain diary. Within-subjects SDs (95%CI) were 11.8 (9.6-15.3), 0.9 (0.7-1.2), and 8.6 (6.9-11.5), respectively. LOA (95%CI) were 32.6 (26.5-42.4), 2.5 (2-3.3), and 23.9 (19.2-31.8), respectively. ICCs (95%CI) were 0.8 (0.6-0.9), 0.8 (0.7-0.9), and 0.9 (0.8-0.9), respectively, but may have been inflated by sample heterogeneity. The test-retest errors allowed detection of clinically relevant effect sizes with feasible sample sizes in future studies, but individual errors were large.

Discussion: For people with PLP, a pain intensity VAS, the SF-MPQ-2, and a pain diary show an acceptable level of inter-session reliability for use in future clinical trials with feasible sample sizes. Nevertheless, the random error observed for all three of the pain outcome measures suggests they should be interpreted with caution in case studies and when monitoring individuals' clinical status and progress.

Key words: reliability; phantom limb pain; outcome measures

Introduction

Phantom limb pain (PLP) is defined as, '… any noxious sensory phenomenon in the missing body part that develops after surgical amputation of a limb' [1]. PLP has a high lifetime prevalence (76%–87%) [2]. PLP has been reported to negatively impact on quality of life for many amputees [3] with intensity of pain the primary issue. A survey of amputees (n = 478), exploring association between pain intensity and interference with activities of daily living, concluded that patients with pain often present with a single primary pain complaint, such as PLP, and seek treatments they hope will decrease pain intensity.

A recent review of PLP indicated a need for the development of standardised quantifiable measures to allow comparison of findings across studies and evaluation of interventions [4]. Clinical trials of pharmacological and non-pharmacological treatments for PLP have predominantly used pain intensity as the primary outcome measure [5, 6], with data collected via patient-reported outcome measures (PROMs). PROMs are particularly relevant, alongside traditional biophysical measures utilised by clinicians, as they capture the patient's subjective experiences of a condition [7, 8]. The visual analogue scale (VAS) for pain intensity and the SF-MPQ-2 have been found to be reliable and sensitive to change in other chronic pain populations [9, 10]. However, there is a complete absence of published studies of the reliability of these outcome measures in people with PLP. Previous findings in other clinical populations cannot be assumed to replicate in PLP. PLP specific studies of these outcome measures are needed to verify their suitability for use with those living with PLP.

Test-retest reliability is a crucial aspect of any outcome measure as it impacts both on the confidence of research inferences and clinical decisions made about individual patients [11]. In a research context, measurements should be reliable enough for a mean treatment effect, at a statistically and clinically meaningful level, to be detected in an intervention trial with appropriate statistical precision and a feasible sample size [12]. In the clinical context, measurements must again be reliable to allow clinical diagnoses and monitoring of status on patients to be robust [13].

The aim of this study was to quantify the test-retest reliability of pain related PROMs in people with PLP. A sub-aim was to assess impact of test-retest errors on future research and clinical decisions.

Materials and Methods

Study overview

A sample of people with PLP consented to undertake three pain outcome measures – the pain intensity VAS [14, 15], the SF-MPQ-2 [10], and a pain diary [16] - at two different time points, with a minimum of seven, and a maximum of 14 days, between measurements. The values at measure one and measure two were compared to assess inter-session (test-retest) reliability. The COSMIN reporting guideline for studies on measurement properties of PROMs was followed in the manuscript [17].

Recruitment

Participants were recruited by consecutive sampling from three NHS hospital trusts in the North-East of England between November 2018 and March 2020. In accordance with guidelines for measurement study sample sizes [13] [18] we aimed to recruit at least 40 participants. Ultimately, 39 participants were recruited for our study. Participants were screened against the eligibility criteria, based on their verbal self-report, by a state registered Occupational Therapist with five years amputee rehabilitation experience.

Inclusion criteria: ≥18 years of age; ≥6 months post amputation; above ankle/wrist amputation; experienced PLP for ≥3 months; experienced PLP rated as ≥3 on a 0-10 scale on at least 2 days in the past week.

Exclusion criteria: Lack of (or any reason to doubt presence of) Mental Capacity to give Informed Consent; currently entered on a research trial investigating a treatment; undergoing active prosthetic rehabilitation e.g., walking training sessions or upper limb Occupational Therapy sessions. The study was approved by Teesside University's School of Health and Life Sciences Research Governance and

Ethics Sub-Committee (ref 084/18) and the East Midlands – Derby Research Ethics Committee (Study number 18/EM/0220), all participants provided written Informed Consent.

PROMs Assessment procedure

At the first contact session an amended version of The Trinity Amputation and Prosthesis Experience Scales Questionnaire-revised (TAPES-R) [19, 20] was used to record the participants' age, sex, ethnicity, medical co-morbidities, amputation history and frequency of PLP (1 -4 [1 = A few times per week, 2 = A few times per day, 3 = A few times per hour, 4= Always]). Participants then completed the first pain intensity VAS and SF-MPQ-2. All data were recorded in the participants home or in a private room at Teesside University. The outcome measures were collected by a state registered Occupational Therapist with five years amputee rehabilitation experience.

Participants placed a mark, representing their average PLP in the previous week, on the VAS - a 100mm line, with anchor statements on the left, "no pain at all" and on the right "worst pain imaginable" [14]. The distance in millimetres to the mark represented the participant's level of PLP intensity for the previous week. Participants then completed the SF-MPQ-2 by rating the extent to which they experienced each of the 22 pain descriptors during the previous week using an 11-point numeric rating scale (0 = "none" to 10 = "worst possible"). The instructions prompted the participant to use 0 if the word did not describe their experience of PLP. The participants were presented with 22 items, each representing a different quality of pain or related symptoms from four summary scales: (1) continuous descriptors (throbbing pain, cramping pain, gnawing pain, aching pain, heavy pain, and tender), (2) intermittent descriptors (shooting pain, stabbing pain, sharp pain, splitting pain, electric-shock pain, and piercing), (3) neuropathic descriptors (hot-burning pain, cold freezing pain, pain caused by light touch, itching, tingling or 'pins and needles', and numbness), and (4) affective descriptors (tiring-exhausting, sickening, fearful, and punishing-cruel). A total PLP score (possible range 0-10) was calculated as the mean rating across all questions [21].

At the end of session one, the first pain diary was given to the participant. Participants completed the diary daily over the next week (seven days) recording a rating of their average PLP intensity for the

day, with reference to a Numerical Rating Scale (0 - 100) where 0 represented no pain at all and 100 represented the worst pain imaginable. The mean of the daily ratings was taken as the overall pain diary score. No imputation was undertaken for missing data – the overall pain diary score was calculated only on available data. If three or more days data were missing the diary was excluded from analysis. Participants also recorded the location of their PLP, anything different or unusual about their pain and the name and dosage of any prescribed or non-prescribed pain medication taken each day, throughout the week. Contact session two occurred 7-14 days after session one. In session two the second VAS and SF-MPQ-2 values were recorded; the second pain diary was issued (to be completed over the next seven days and returned).

Statistical analysis

Several analysis approaches were used for quantifying and interpretating test-retest reliability in keeping with the different contexts of measurement; these being research on a sample of patients and for assessment of individual patient status [13]. The systematic (mean change) and random errors between session one and two for the three PROMs were quantified. The fundamental issues are whether the test-retest variability is small enough to allow robust decisions to be made when the measurement tool is used in future research or to help arrive at robust clinical decisions on individual patients [13]. Histograms and Q–Q plots were used to assess the distribution of the data of the differences (residuals) between session one and two [22]. A paired t-test was used to quantify the mean (95% CI) change between session one and two, i.e., to explore any evidence of systematic changes (bias). Random error was quantified with the within-subjects SD, the coefficient of variation (CV), 95% limits of agreement (LOA) and a two-way mixed effects model intraclass correlation coefficient (ICC). Point estimates of each error statistic were presented alongside the respective 95% confidence intervals (95% CI). When the 95% CI of ICC is <0.5 reliability is considered poor, between 0.5 to 0.75 moderate, between 0.75 to 0.90 good, and greater than 0.90 excellent [23]. Nevertheless, it is important to also gauge more direct “analytical goals” such as individual diagnostic robustness and impact on future research in reliability studies [13], which is discussed in Supplementary file 1. Bland-Altman plots [24] were drawn for each outcome measure to visually

assess the mean difference (systematic error), data scatter and the relationship between magnitude of difference and size of measurement. Evidence of heteroscedasticity can be observed when a left-to-right trend of increasing scatter exists in the differences, in a "shotgun" manner [24]. The LOA are useful for interpreting the amount of test-retest error expected for an individual patient. We also estimated whether the test-retest error was small enough for typical and clinically relevant effect sizes to be detected as statistically significant in future studies with feasible sample sizes [13]. For this process, we selected a minimal clinical important difference (MCID), as the smallest effect that is of clinical significance and is a key parameter in sample size estimation [12]. Currently there are no established MCIDs for people with PLP, however, previous NICE guidelines have agreed on a consensus MCID between groups of a 10% change in measurement for pain [25, 26]. Based on this, the MCIDs for the three pain PROMs were 1/10 (SF-MPQ-2) and 10/100 (VAS and pain diary). Given these effect sizes, the SD of the test-retest differences was then used to estimate sample sizes for future studies using a general stand-alone power analysis program for statistical tests - G*Power 3.1 [27]. All data were analysed, and Bland-Altman plots drawn, using Microsoft Excel and the IBM SPSS Statistics package (version 26).

Results

Participant characteristics

Forty-eight people were screened for eligibility, five were excluded (one was less than 6 months post amputation, one had a single digit amputation and three had infection in their residual limb), three decided not to take part and 40 provided Informed Consent. One, was withdrawn following the first session due to symptoms of a previous ischemic stroke recrudescence during the testing period. Thirty-nine (30/9 males/females, mean [SD] age 55.3 [16.2], and mean [SD] of 6.8 [8.3] years post amputation) took part. The frequency (median/IQR) of PLP was 1/1 (a few times per week) at baseline and stable between time points. Nineteen of the thirty-nine participants recorded their pain medication. This was stable between time points. Thirty-seven participants (28/9 males/females, mean [SD] age 57 [14], 36 lower and 1 upper limb amputee) provided complete data sets for VAS and SF-MPQ-2. Thirty-two participants (mean [SD] age 56 [14], 23/9 males/females, 22 lower and 1 upper

limb amputee) for pain diaries (Figure 1). Summary statistics for the recorded values of the two measurements of each PROM, are shown in Table 1.

Figure 1 near here

Legend: Figure 1. Flow chart of participant recruitment process, testing and data analysis.

Table 1 near here

Test-retest reliability statistics

The test–retest reliability statistics are shown in Table 2 and the three Bland–Altman plots are included in Figure 2. The test-retest differences (residuals) were found to be reasonably normally distributed. Data from thirty-seven participants were analysed for the VAS and SF-MPQ-2 and thirty-two participants were analysed for the pain diary.

Pain VAS

The mean (SD) for session one of the VAS was 58.1 (24.9) and for session two was 58.7 (24.3). The difference between session 1 and 2 was negligible (i.e., only 6% of our proposed MCID) and not statistically significant ($p = 0.8$), indicating no evidence of systematic mean change between test and retest. The mean inter-session difference indicated no trend towards scoring higher or lower on the second session, inclusive of the 95% CI, thus there was no evidence of order effects. The ICC value for absolute agreement was 0.8 (95% CI: 0.6 – 0.9), indicating moderate to good level of reliability, although such test-retest correlation statistics can be inflated by sample heterogeneity. Visual inspection of the Bland-Altman plot showed no evidence of heteroscedasticity or proportional bias. The limits of agreement (LOA) are wide indicating a relatively large amount of individual patient random error between measurements 1 and 2.

SF-MPQ-2

The mean (SD) for session one of the SF-MPQ-2 was 3.4 (2.2) and for session two was 3.3 (2.1). The difference between session 1 and 2 was negligible (i.e., only 1% of the MCID) and not statistically significant ($p = 0.7$), thus there was no evidence of systematic bias. The 95% CI of the mean session difference showed no trend towards scoring higher or lower on the second session, thus there was no evidence of order effects. The ICC value for absolute agreement was 0.8 (95% CI: 0.7 – 0.9), indicating moderate to good level of reliability. Visual inspection of the Bland-Altman plot shows homoscedastic errors across the measurement range and thus no evidence of heteroscedasticity or

proportional bias. The wide LOA relative to the MCID indicates a large amount of random error for the test-retest reliability of this PROM.

Pain Diary

The mean (SD) for session one of the pain diaries was 49.3 (23.8) and for session two was 47.1 (25.3). The difference between session 1 and 2 was small (i.e., 22% of the MCID) and not statistically significant ($p = 0.3$), thus there was no evidence of systematic bias. The 95% CI of the mean session difference showed a slight trend towards scoring higher on the second diary, but the magnitude of change was below the MCID. The ICC value for absolute agreement was 0.9 (95% CI: 0.8 – 0.9), indicating good to excellent level of reliability. There was no evidence of heteroscedasticity. The LOA were wide relative to the MCID indicating large amounts of random error on an individual patient basis.

Table 2 near here

Figure 2 near here. Legend: Figure 2 Bland-Altman plots for the three pain PROMs. The differences between the tests are plotted against each individual's mean for the two sessions. Mean session difference (systematic bias) is displayed by a solid line and limits of agreement by dashed lines.

Sample size estimations for future research

Based upon the SD of the differences obtained from the test-retest design, and using an alpha level of 0.05, and a power of 90%, in a future two-armed RCT, to detect a change equivalent to the MCID, sample sizes of 66, 70 and 120 participants would be required for the pain diary, SFMPQ-2, and the pain VAS, respectively. In a single arm pre/post study the sample of participants would be 18, 19 and 32 for the pain diary, SFMPQ-2 and the pain VAS, respectively.

Discussion

The aim of this study was to investigate the inter-session reliability of three pain PROMS for people living with PLP. The test-retest reliability of each: Visual Analogue Scale (VAS) for pain intensity, SF-MPQ-2, and a pain diary was quantified over two data collections. Any systematic changes were

quantified by consideration of the mean difference and associated 95% CI between sessions. The mean differences between session 1 and 2 did not reach statistical significance for any of the three pain measures. Furthermore, the mean test-retest changes for all three of the pain related PROMs were less than our selected MCID and this was inclusive of the associated 95% confidence intervals. Therefore, we could not detect any clinically relevant order effects (e.g., learning effect) across all three PROMs over a seven-to-fourteen-day period.

With regards to random error, the within-subjects SD (the standard error of measurement [SEM] is calculated in the same way) is a useful statistic as it relates to true score theory and is expressed in the actual units of measurement and can be compared between outcome measures in relation to their MCID [28]. In terms of SEM the pain diary was the most reliable of the pain outcome measures, demonstrating a SEM of 8.6 points in comparison to the SF-MPQ-2, with a SEM of 0.9 (MCID 1/10) and the pain VAS, with a SEM of 11.8 points. The SEM is below the MCID (14% and 10% for the pain diary and SF-MPQ-2 respectively) and slightly above the MCID (18% for the pain intensity VAS) and this increases confidence in being able to distinguish true change in the condition from that which has occurred from the measurement error associated with the outcome measures [13]. Higher levels of random error were identified in comparison to previous test-retest reliability studies of chronic pain populations, in which SEM was calculated. For the osteoarthritic knee pain population, the SEM was 0.3 for the VAS and 4.8 for the NRS [29] compared to 11.8 (9.6 - 15.3) and 8.6 (6.9 - 11.5) in our study population respectively. For the musculoskeletal shoulder pain population, the SEM of the SF-MPQ-2 was 0.5 [30] compared to 0.9 (0.7 - 1.2) in our study population. It could be suggested that the episodic nature of PLP contributes to the higher levels of random error found in our study population. On an individual level, the limits of agreement (LOA) give an indication of the upper magnitude of how much it is expected (95%) that an individual could potentially change in status because of natural variation from measurement to measurement [24]. Considering each of the PROMs studies the upper LOAs (as seen in the Bland-Altman plots - Figure 2) a slightly different picture emerges with LOA for the pain diary ±24/100, the SF-MPQ-2 ±2.5/10 and ±33/100 for the pain intensity VAS. The pain related PROMs contain a random error component around two and half

times the MCID in the pain diary and SF-MPQ-2 and over three times in the pain VAS. Within this population there is a large degree of random error relative to the MCID thus it is questionable whether the measure is sufficiently reliable for clinical purposes in individual patients. This implies that monitoring change on an individual patient level should be done with caution so that apparent change in the outcome is not misinterpreted beyond the natural variability of the outcome measure.

Recall bias may have accounted for the slightly smaller measurement error for the pain diary as this was completed daily in comparison to the pain VAS which required a weekly recall of average pain intensity. It could be that completion prospectively over seven days potentially increases the participant's accuracy of recording a condition which is episodic and varies in intensity by its very nature [16]. Previous large scale (n= 255) cross-sectional survey research [31] has identified that assessing PLP severity can be vulnerable to recall bias because PLP is on average experienced as brief and episodic, not constant. This may underscore the importance of averaging multiple measures of pain intensity across time to maximize reliability as previously reported in a few studies that have investigated the psychometric strengths of composite NRS scores in chronic pain intensity assessment [32-34].

VAS ratings (100mm line) recorded in this study were based on the participant recalling an average intensity score over the preceding week. Future studies looking at treatment effects in the PLP population may benefit from utilizing a composite score that incorporates a rating of pain intensity with the frequency that it was experienced during the week. This may provide a score that accounts for the individual daily fluctuations in PLP that contributes to the random variability in pain intensity scores (from the NRS pain diaries) evident in this study. Dividing intensity (0-100) by frequency (1 = all the time to 5 = once or less per month), referred to as 'Chronic PLP', has been found to be reliable [35] and has been used as a primary outcome measure in recent fMRI studies with PLP participants [36-39]. The SF-MPQ-2 is also a composite score (of pain dimensions - continuous, intermittent, neuropathic, and affective) but unlike the pain diary the scores are collected on one occasion, making it less burdensome for the participant to complete. The reliability results were similar between the two

outcome measures, but the SF-MPQ-2 may be more feasible for the research setting considering the higher number of non-returned pain diaries.

Strengths and limitations

A strength of the study is that it is the first to assess the inter-session reliability of PROMS for people with PLP, that are commonly used in trials involving PLP participants. This provides reliability data that are more generalisable to this clinical population than has been previously published. As previously published reliability studies have not included people with PLP. However, the sample included predominately male participants with a lower limb amputation and therefore the findings may be less generalisable to females or people with an upper limb amputation. The sample size for the study is below that recommended by the COSMIN guidelines [17]. However, the sample size for the VAS and SF-MPQ-2 is in keeping with those recommended by Bland and Altman [24] to ensure adequate precision of estimates However, seven participants did not return their first, second or both pain diaries preventing their inclusion in the test-retest reliability assessment. The 26% non-returned diary rate was a limitation of the study in terms of adherence to the protocol. Furthermore, this reduces the precision of measurement error estimates for this PROM. Given this smaller sample size, the associated 95% confidence intervals around the random error statistics (e.g., SEM and LOA) should be considered as a less precise estimation of measurement error than would have been achieved with a larger sample.

Future research and Clinical implications

These findings indicate that the SF-MPQ-2 and the pain diary could be classed as sufficiently reliable for research purposes. The estimated sample sizes required for an RCT are achievable and within normal expectations, based on previous RCTs that have tested therapeutic interventions for PLP [40, 41]. The pain VAS should be treated with caution as it may be underpowered to detect a statistically significant and clinically worthwhile change, if the estimated sample size (n = 120) is not achieved. Multi-site recruitment may be required to achieve such an ambitious target for the PLP population. Future research could consider the test-retest reliability of a chronicity based composite score of PLP

that incorporates intensity, frequency, and duration of the experience. That may prove a reliable - and potentially more valid - measurement of PLP as it may better capture patient experience of what is an episodic condition, with flare-ups of differing durations, occurring on a regular basis.

Overall, the SF-MPQ-2 may be preferable as the primary outcome measure for research trials given the similar sample size estimations to the pain diary but demonstrated better completion rates and less participant burden. However, the limits of agreements were relatively wide for each measure compared to their respective MCIDs; thus, clinicians should interpret these measures with a degree of caution at the individual patient level within clinical practice.

Conclusion

This study presented data on the test-retest reliability of the VAS, SF-MPQ-2, and a pain diary in people with PLP. The data suggests that the three pain specific PROMs have an acceptable amount of measurement error for practical use in research trials. However, the quantified amounts of random error in all three of the pain PROMs suggest that they should be used with caution for case studies or the assessment and monitoring of individual patients.

References

1. Schug SA, Lavand'homme A, Barke B, et al. The IASP classification of chronic pain for ICD-11: chronic postsurgical or posttraumatic pain. *Pain*, 2019; 160; 45-52.

2. Stankevicius A, Wallwork SJ, Summers B, et al. Prevalence and Incidence of Phantom Limb Pain, Phantom Limb Sensations and Telescoping in Amputees: A Systematic Rapid Review. *Eur J Pain*, 2020; 25; 23-28.

3. Van der Schans CP, Geertzen J HB, Schoppen T, et al. Phantom pain and health-related quality of life in lower limb amputees. *J Pain Symp Mgt*. 2002; 24; 429-436.

4. Schone HR, Baker CI, Katz J, et al. Making sense of phantom limb pain. Journal of Neurology. *Neurosurgery & Psychiatry*. 2022; 93; 833-843.

5. Batsford S, Ryan CG, and Martin DJ. Non-pharmacological conservative therapy for phantom limb pain: A systematic review of randomized controlled trials. *Physiother Theory Pract*. 2017; 33; 173-183.

6. Alviar M, Hale T, and Dungca M. Pharmacologic interventions for treating phantom limb pain. *Cochrane Database Syst Rev*. 2016; 10; CD006380.

7. Kyte DG, Calvert M, Van der Wees PJ, et al. An introduction to patient-reported outcome measures (PROMs) in physiotherapy. *Physiotherapy*. 2015; 101; 2; 119-25.

8. Kyte D, IJ, Draper D, et al. Current practices in patient-reported outcome (PRO) data collection in clinical trials: a cross-sectional survey of UK trial staff and management. *BMJ Open*. 2016; 6; p1-13

9. Ferraz MB, Quaresma MR, Aquino LR, et al. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. *J of Rheumatology*. 1990; 17; 1022-1024.

10. Dworkin RH, Turk DC, Revicki D A, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). *Pain*. 2009; 144; 35-42.

11. Atkinson, G and Batterham AM. True and false interindividual differences in the physiological response to an intervention. *Experimental Physiology*. 2015; 100; 577-588.

12. Batterham AM and Atkinson G. How big does my sample need to be? A primer on the murky world of sample size estimation. *Phys Therapy in Sport*. 2005; 6; 153-163.

13. Atkinson G and Nevill AM. Statistical Methods for Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine. *Sports Medicine.* 1998. 26; 217-238.

14. Scott J and Huskisson E. Graphic representation of pain. *Pain*. 1976; 2; 175-184.

15. Carlsson AM. Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. *Pain*. 1983;16; 87-101.

16. Whyte AS and Niven CA. Variation in phantom limb pain: results of a diary study. *J of Pain and Symp Management*. 2001; 22; 947-953.

17. Gagnier JJ, Lai J, Mokkink LB, et al. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Quality of Life Research,* 2021; 30; 2197-2218.

18. Altman, DG. *Practical statistics for medical research*. Boca Raton: CRC press, 1990.

19. Gallagher P, Franchignoni F, Giordano A, et al. Trinity amputation and prosthesis experience scales: a psychometric assessment using classical test theory and rasch analysis. *American J of Phys Med & Rehab*. 2010; 89; 487-496.

20. Gallagher P, MacLachlan M. Development and psychometric evaluation of the Trinity Amputation and Prosthesis Experience Scales (TAPES). *Rehab Psychology*. 2000; 45; 130-154.

21. Dworkin RH, Turk DC, Revicki, Dennis A, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). *Pain*; 2009; 144; 35-42.

22. Field, A. *Discovering statistics using IBM SPSS statistics*. 5th ed. Newbury Park: Sage, 2018.

23. Koo TK and Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiro Med*. 2016; 15; 155-163.

24. Bland JM and Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Med Res*.1999; 8; 135-160.

25. National Institute for Health and Care Excellence (NICE). Draft: Low back pain and sciatica: management of non-specific low back pain and sciatica. *Assessment and noninvasive treatments. 2016. Available from: https://www.nice.org.uk/guidance/GIDCGWAVE0681/documents/draft-guideline, accessed 25th June 2021.*

26. National Institute for Health and Care Excellence (NICE). Evidence review for electrical physical modalities for chronic primary pain: Chronic pain (primary and secondary) in over 16s: assessment of all chronic pain and management of chronic primary pain: Evidence review H. Available from https://www.nice.org.uk/guidance/ng193/evidence/h-electrical-physical-modalities-for-chronic-primary-pain-pdf-9071987013, Accessed 25th June 2021.

27. Faul F, Erdfelder E, Lang AG, et al. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Res Methods.* 2007. 39; 175-191.

28. Hopkins, WG. Measures of reliability in sports medicine and science. *Sports Med*. 2000; 30; 1-15.

29. Alghadir AH, Anwer S, Iqbal A, et al. Test–retest reliability, validity, and minimum detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. *J Pain Res.* 2018; 851-856.

30. Jumbo SU, MacDermid JC, Packham TL, et al. Reproducibility: reliability and agreement parameters of the Revised Short McGill Pain Questionnaire Version-2 for use in patients with musculoskeletal shoulder pain. *Health & QoL Outcomes*. 2020; 18; 1-12.

31. Ehde DM, Czerniecki JM, Smith DG, et al. Chronic phantom sensations, phantom pain, residual limb pain, and other regional pain after lower limb amputation. *Arch Phys Med & Rehab*. 2000; 81; 1039-1044.

32. Jensen MP and McFarland CM. Increasing the reliability and validity of pain intensity measurement in chronic pain patients. *Pain.* 1993; 55; 195-203.

33. Jensen, M.P., et al., Comparative reliability and validity of chronic pain intensity measures. *Pain*. 1999; 83; 157-162.

34. Jensen MP, Turner JA, Romano JM, et al. The use of multiple-item scales for pain intensity measurement in chronic pain patients. *Pain.* 1996; 67; 35-40.

35. Kikkert S, Mezue M, Slater DH, et al. Motor correlates of phantom limb pain. *Cortex*. 2017; 95; 29-36.

36. Kikkert S, Johansen-Berg H, Tracey I, et al. Reaffirming the link between chronic phantom limb pain and maintained missing hand representation. *Cortex.* 2018, 106; 174-184.

37. Makin, T.R., et al., *Network-level reorganisation of functional connectivity following arm amputation.* Neuroimage, 2015. 114: p. 217-25.

38. Makin TR, Filippini N, Duff EP, et al. Phantom pain is associated with preserved structure and function in the former hand area. *Nat Commun*. 2013; 4; 1570.

39. Wesselink DB, Dempsey-Jones H, Cardinali L, et al. Obtaining and maintaining cortical hand representation as evidenced from acquired and congenital handlessness. *E-life*. 2019; 8; 37227.

40. Rothgangel A, Braun S, Winkens B, et al. Traditional and augmented reality mirror therapy for patients with chronic phantom limb pain (PACT study): results of a three-group, multicentre single-blind randomized controlled trial. Clin Rehab. 2018; 32; 1591-1608.

41. Gunduz ME, Pacheco-Barrios K, Bonin Pinto C, et al. Effects of Combined and Alone Transcranial Motor Cortex Stimulation and Mirror Therapy in Phantom Limb Pain: A Randomized Factorial Trial. *Neurorehab & Neural Repair*. 2021; 35; 704-16.

Figure 1

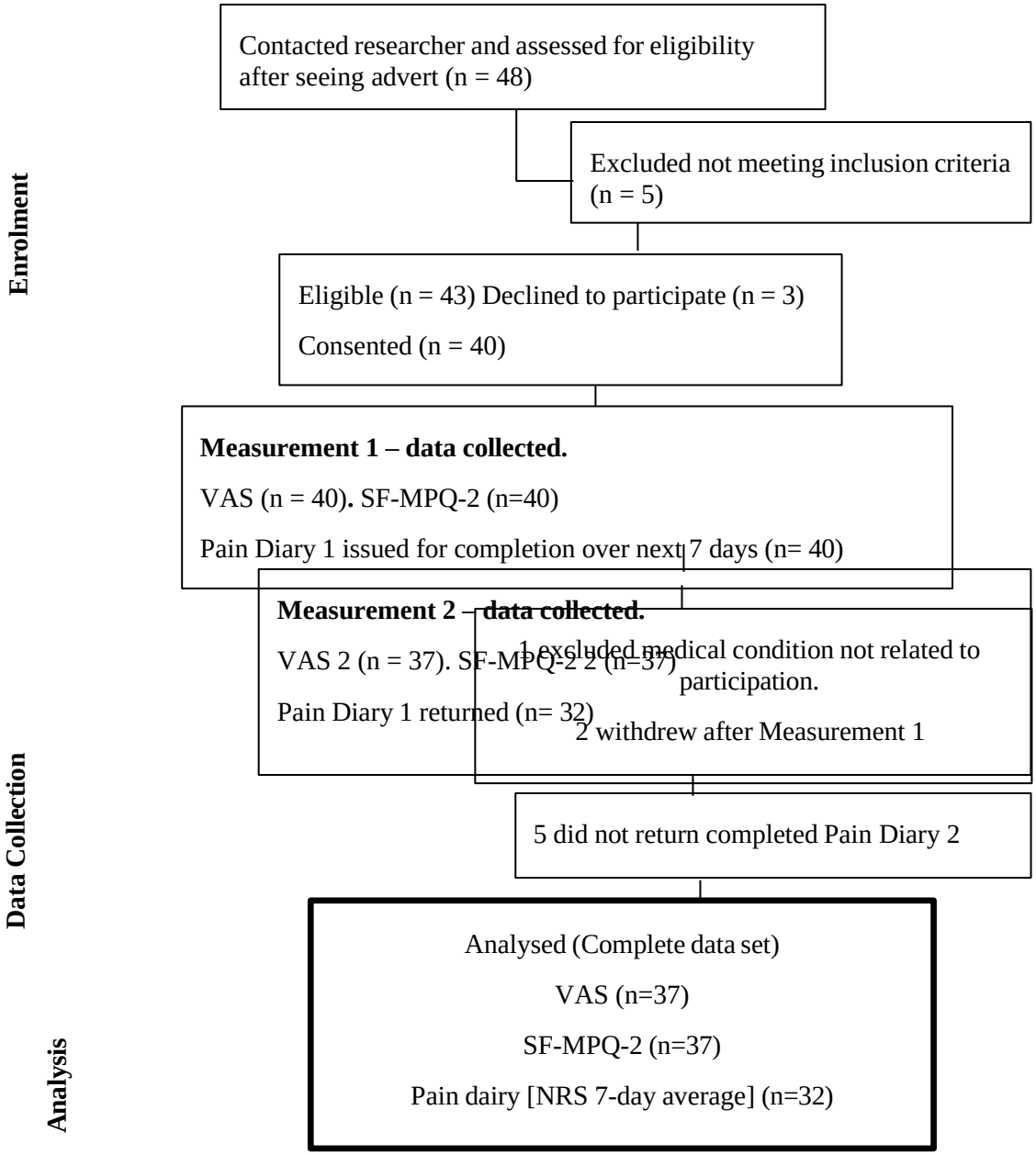


Figure 1. Flow chart of participant recruitment process, testing and data analysis.

Table 1 PROMS summary statistics - mean (SD)

PROM	**Measurement 1**	**Measurement 2**	**Difference – Mean (95% CI)** **SD (95% CI)**
VAS (0-100)	58.1 (24.9)	58.7 (24.3)	+0.6 (-6.1 - 5) 16.6 (13.5 - 21.6)
SF-MPQ-2 (0-10)	3.4 (2.2)	3.3 (2.1)	-0.1 (-0.3 - 0.5) 1.3 (1.0 - 1.7)
Pain Diary NRS (0-100)	49.3 (23.8)	47.1 (25.3)	- 2.2 (-2.2 - 6.6) 12.2 (9.8 - 16.2)

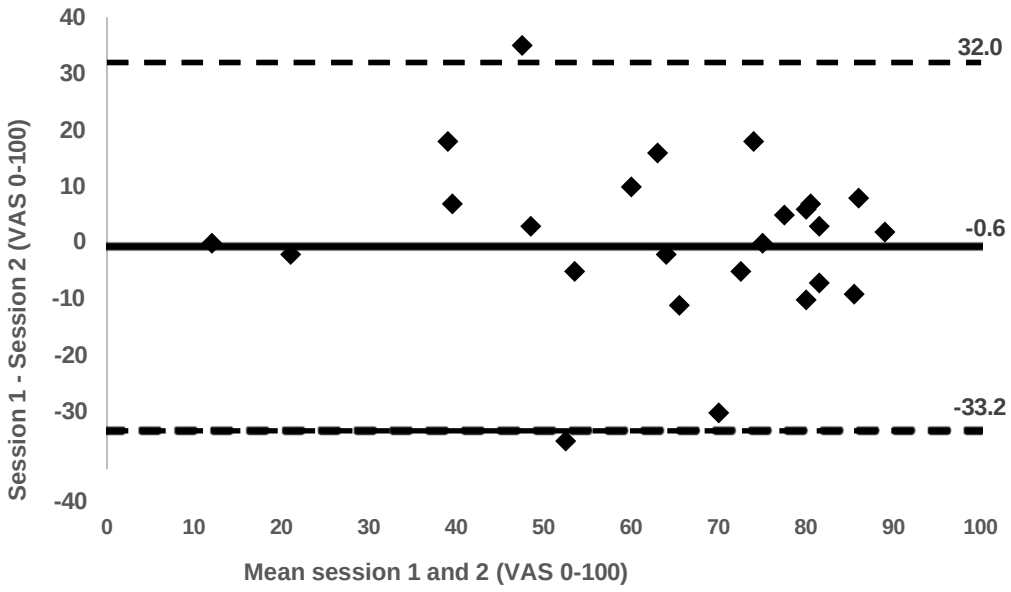
VAS = Visual Analogue Scale; SF-MPQ-2 = Short-form McGill Pain Questionnaire; NRS = Numeric Rating Scale

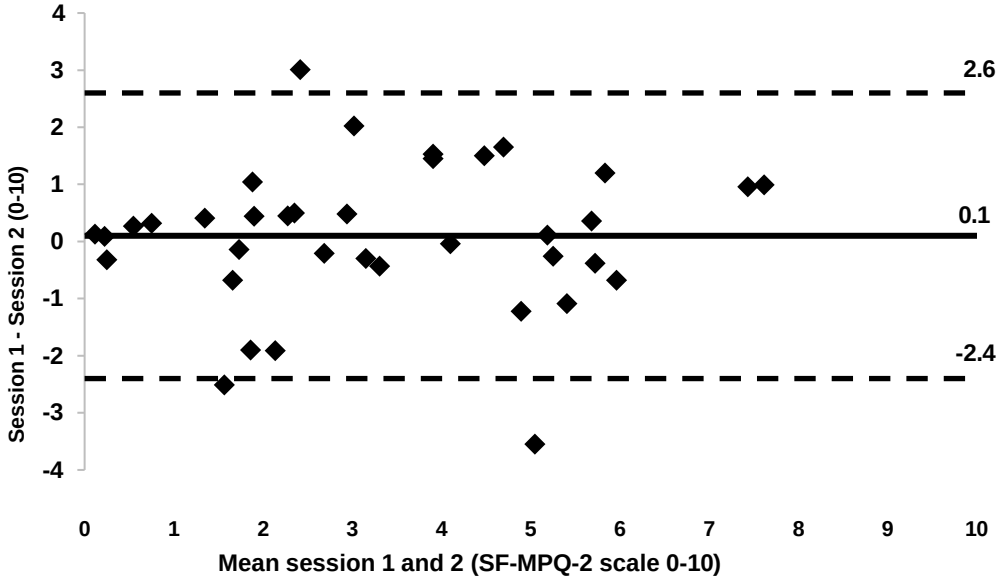
Table 2 Reliability statistics for pain outcome measures

	VAS	**SF-MPQ-2**	**Pain diary**
Mean session difference (95% CI)	-0.6 (-6.1 - 5)	0.1 (-0.3 - 0.5)	2.2 (-2.2 - 6.6)
SD of session differences (95% CI)	16.6 (13.5 - 21.6)	1.3 (1.0 - 1.7)	12.2 (9.8 - 16.2)
Within-subjects SD (SEM) (95% CI)	11.8 (9.6 - 15.3)	0.9 (0.7 - 1.2)	8.6 (6.9 - 11.5)
Coefficient of variation (%) (95% CI)	20.1(16.4 - 26.2)	27.1 (22.1 - 35.2)	17.9 (14.3 - 23.8)
Limits of agreement (95% CI)	32.6 (26.5 - 42.4)	2.50 (2 - 3.3)	23.9 (19.1 - 31.8)
ICC (95% CI)	0.8 (0.6 - 0.9)	0.8 (0.7 - 0.9)	0.9 (0.8 - 0.9)

SD = Standard Deviation; SEM = standard error of measurement (typical error); $ICC^{3.1}$ = Intraclass correlation coefficient; CI = confidence interval.

Figure 2


40
30
20
10
0
-10
-20
-30
-40
32.0
-0.6
-33.2
Session 1 - Session 2 (VAS 0-100)
0
10
20
30
40
50
60
70
80
90
100
Mean session 1 and 2 (VAS 0-100)


4
3
2
1
0
-1
-2
-3
-4
2.6
0.1
-2.4
Session 1 - Session 2 (0-10)
0
1
2
3
4
5
6
7
8
9
10
Mean session 1 and 2 (SF-MPQ-2 scale 0-10)

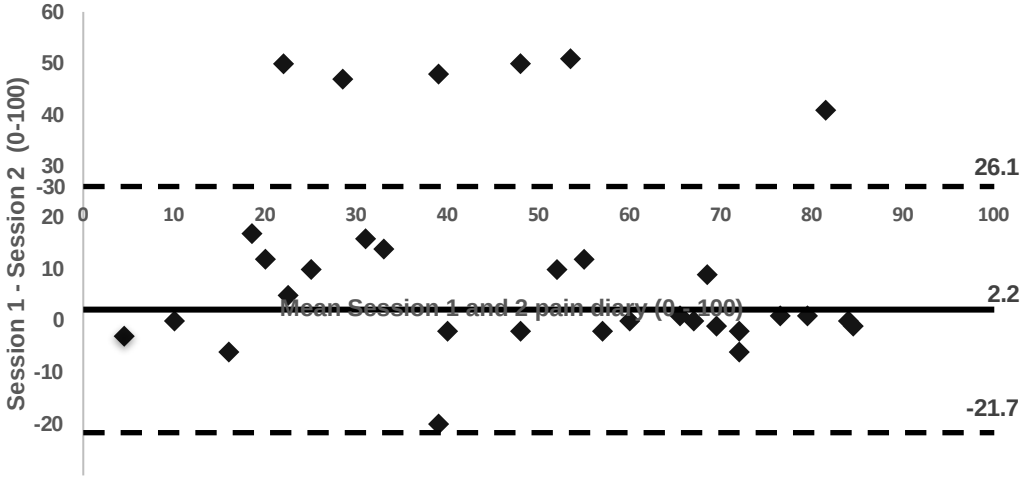

60
50
40
30
20
10
0
-10
-20
26.1
2.2
-21.7
Session 1 - Session 2 (0-100)
0
10
20
30
40
50
60
70
80
90
100
Mean Session 1 and 2 pain diary (0 - 100)

Figure 2 Bland-Altman plots for the three pain PROMs

The differences between the tests are plotted against each individual's mean for the two sessions. Mean session difference (systematic bias) is displayed by a solid line and limits of agreement by dashed lines.