Comparison of reliability, construct validity and responsiveness of the IPAQ-SF and PASE in adults with osteoarthritis.

Abstract

Introduction

This study assessed the measurement properties of two commonly used self-report physical activity (PA) measures: the International Physical Activity Questionnaire- Short Form (IPAQ-SF) and the Physical Activity Scale for the elderly (PASE) in adults with osteoarthritis.

Methods

Secondary analysis of the MOSAICS cluster randomised controlled trial baseline and 3-month follow up questionnaires; total scores and subdomains of the IPAQ-SF and PASE were compared. Intra-class correlations (ICC) were used to assess test-retest reliability, measurement error was assessed using standard error of measurement (SEM), smallest detectable change (SDC) and 95% limits of agreement (LoA). Responsiveness was assessed using effect size (ES), standard responsive measurement (SRM) and response ratio (RR).

Results

There was moderate correlation (r=0.56) between the total IPAQ-SF scores (score ranges 0-16398) and the total PASE scores (score ranges 0-400). Subdomain correlations were also moderate (ranges 0.39-0.57). The PASE showed greater reliability compared to the IPAQ-SF (ICC=0.68; 0.61-0.74 95%CI & ICC=0.64; 0.55-0.72, respectively). Measurement error in both measures were large: PASE SEM=46.7, SDC=129.6 and 95% LoA ranges=-117 to 136, the IPAQ-SF SEM=3532.2 METS$^{-1mins-1week}$, SDC= 9790.8 and 95% LoA ranges=-5222 to 5597. Responsiveness was poor: ES -0.14 and -0.16, SRM -0.21 and -0.21 and RR 0.12 and 0.09 for the IPAQ-SF and PASE respectively.

Discussion

The IPAQ-SF and PASE appear limited in reliability, measurement error and responsiveness. Researchers and clinicians should be aware of these limitations, particularly when comparing different levels of PA and monitoring PA levels changes over time in those with osteoarthritis.

Keywords

Measurement properties, osteoarthritis, physical activity, IPAQ, PASE

Introduction

Osteoarthritis (OA) is the most common cause of peripheral joint pain in adults aged 45 years and over (Felson, 2009) . It is a clinical syndrome of joint pain causing varying degrees of limitation in physical function and reduced quality of life (National Institute for Health and Care Excellence (NICE) 2014). Physical activity (PA) interventions are effective in reducing pain and improving physical functioning among those with lower limb OA (Fransen et al., 2015; Fransen, McConnell, Hernandez-Molina, & Reichenbach, 2014; Holden et al., 2015; Uthman et al., 2013). In addition, a physically active lifestyle has wider health benefits reducing the risk of premature mortality, disability, chronic diseases and mental health conditions (Warburton, Nicol, & Bredin, 2006). Due to its benefits, PA is recommended as a core treatment for all adults with OA (NICE, 2014). However, current levels of physical activity in people with OA are low (Herbolsheimer et al., 2016).

Despite PA recommendations, it is not clear what frequency, duration or intensity of PA is required for those with OA to gain clinically important benefits in pain and physical functioning (Quicke, Foster, Croft, Ogollah, & Holden, 2018). In order to draw inferences regarding these parameters of PA, self-report measures also need to accurately capture these elements of PA. Valid and reliable self-report measures of PA are required to establish current PA levels in people with OA, whilst responsive measures are necessary to detect changes over time or following interventions.

There are two main approaches available for measuring levels of PA: objective measures (doubly labelled water, indirect calorimetry or activity monitors such as pedometers and accelerometers) and subjective measures (self-report questionnaires and activity diaries). Self-report measures of PA are an attractive approach as they are inexpensive to administer, have the potential to measure all forms of PA, can be self-completed and so used in large population level studies (Prince et al., 2008). The limitations of self-report measures are that they can both over and underestimate levels of PA due to reporting bias, recall bias and social desirability bias which may affect the reliability and validity of their measurement of PA. Poor validity has been most notably identified at higher intensities of PA where greater differences were found between self-report and objective measures (Prince et al., 2008; Silsbury, Goldsmith, & Rushton, 2015).

Self-report PA measures are commonly used in OA research. A recent systematic review showed that self-report measures of PA have been used in 91 studies since 2018 (Smith et al., 2019). The International Physical Activity Questionnaire - Short Form (IPAQ-SF) and the Physical Activity Scales for the Elderly (PASE) have been most commonly used. However, recent reviews identified a lack of evidence for the validity and reliability of these measures in adults with OA (Healey et al., 2020; Smith et al., 2019). Importantly, to date, neither the IPAQ-SF nor PASE have been assessed for responsiveness (ability to detect changes in levels of PA over time). The aim of this study was to evaluate and compare the test-retest reliability, construct validity and responsiveness of the IPAQ-SF and PASE in adults with OA in the hand, hip, knee or foot.

Methods

Design

This study was conducted using secondary analysis of data from a cluster-randomised controlled trial (RCT): Managing OSteoArthritis In ConsultationS (MOSAICS) study (ISRCTN number: ISRCTN06984617). The protocol and results of the MOSAICS study are reported elsewhere (Dziedzic et al., 2018; Dziedzic et al., 2014). Briefly, the MOSAICS study was a two-arm cluster Randomised Controlled Trial (RCT) conducted in eight UK general practices. Adults aged 45 years and over, who were registered with participating general practices were mailed a health survey. Participants that reported peripheral joint pain who consented to follow-up and consulted their General Practitioner (GP) for joint pain were invited to take part in the cluster RCT. Participants in the intervention practices received an enhanced GP consultation, an OA self-management guidebook and were offered follow-up practice nurse consultations, where core-recommended treatments were delivered (Grime & Dudley, 2014). Participants in the control arm practices received usual care.

Participants

All 525 participants from the MOSAICS study cluster trial (288 from intervention practices and 237 from control practices), were included in this study. Inclusion criteria for the MOSAICS study were adults aged 45 and over, consulting with joint pain in the hand, hip, knee or foot, at a participating practice. Exclusion criteria comprised those who were: screened as ineligible by GP screening of practice list; unable to give consent; a resident in a nursing home; had a history of serious disease (malignancy, terminal illness), unable to consult to their GP or flagged on the practice list as excluded from research (Dziedzic et al., 2014). Data from the baseline and 3-month follow-up of the MOSAICS questionnaires were used for this study.

Physical activity measures

The PASE

The PASE is designed specifically to measure levels of PA in adults aged 65 and over. The PASE gives an output score ranging from 0-400. The scoring of the PASE does not represent a quantifiable amount of activity time which could indicate if a participant is participating in low, moderate or high amounts of physical activity or meeting guideline-recommended activity levels, instead higher scores represent higher levels of PA. The PASE is a short self-report measure with items on PA in leisure, occupational and household settings (Washburn, Smith, Jette, & Janney, 1993). The measure was scored according to the instrument scoring guideline. Outliers of the PASE were checked if individuals' total score exceeded 400 as recommended within the scoring guideline.

The IPAQ-SF

The IPAQ-SF was developed as an outcome to measure levels of PA for comparisons across international populations. The IPAQ-SF measures energy expenditure per week (METS$^{-1mins-1week}$) and can give a continuous or categorical score rating of an individual's weekly PA level; low (<600 METS$^{-1mins-1week}$), moderate (≥600-2999 METS$^{-1mins-1week}$) and high PA levels (≥3000 METS$^{-1mins-1week}$). The IPAQ-SF contains four items which assess sedentary activities, walking activities, moderate intensity activities and vigorous intensity activities (Booth et al., 2003). The measure was scored according to the instrument scoring guideline and was conducted for both the categorical output and continuous output. While there is not an upper limit on the scoring range of the IPAQ-SF; truncation of the IPAQ-SF data was conducted in participants that scored higher than a total of 3 hours of either walking, moderate or vigorous activities per day, to allow a maximum of 3 hours per day in each activity. Outliers that where the sum of walking, moderate and vigorous activity totalling at greater than 21 hours were excluded as recommended in the IPAQ-SF scoring manual.

Muscle strengthening exercises or general fitness exercise

Uptake of NICE core exercise recommendations was measured using a previously validated questionnaire (Jinks, Jordan, Ong, & Croft, 2003; NICE, 2014). Participants were asked to report if they had tried muscle strengthening exercises for their joint pain or general aerobic fitness exercise for their joint pain in the last three months.

Participant characteristics

Self-reported participant characteristics and longitudinal descriptive statistics (baseline and three months) were collected for this study. Participant baseline characteristics included age and gender. Longitudinal descriptive statistics included: body mass index (BMI, kg/m^2), pain intensity (calculated using a 0-10 numerical rating scale for each peripheral joint site, for those with multiple sites of pain we took the score from the joint with the highest rated pain (Finney, Dziedzic, Lewis, & Healey, 2017)), health status was measured using the SF-12 (Ware Jr, Kosinski, & Keller, 1996), both the physical component scale (PCS) and mental component scale (MCS) of the SF-12 were used, with a score range of 0-100 with lower score indicating lower levels of health, and quality of life (QoL) (measured using the EQ-5D with a score range from 0-1 with lower score indicating lower QoL (3-level response) (EuroQol Group, 1990)). At 3 months follow-up participants recorded a global assessment of change from baseline with ranges from 1= completely recovered to 6= much worse (K. S. Dziedzic et al., 2014).

Procedure

Reliability & measurement error sub-sample

Reliability and measurement error of the IPAQ-SF and PASE were assessed between baseline and 3-month follow-up in a sub-group of participants who appeared to have remained stable in terms of their PA level during the study period. The reliability sub-group of participants included: those who completed either the IPAQ-SF or PASE at baseline and 3-month follow-up and self-reported no change in their self-reported physical activity behaviour questions from baseline to follow-up. For example, reported not trying muscle strengthening exercises or general fitness exercises at both baseline and 3-month follow-up, or reported trying muscle strengthening exercises or general fitness exercises at both baseline and three months follow-up. A 3-month follow-up period was selected as an appropriate second measurement time-period for evaluating the reliability and measurement error of the IPAQ-SF /PASE as most RCT investigating PA interventions for people with lower limb OA are of 2-3 months in duration (Juhl et al., 2014). It is of clinical importance to understand the test-retest reliability of the PASE and IPAQ SF over a similar time-period. This knowledge would help researchers and clinicians to understand whether changes in scores are due to real change or measurement error.

Responsiveness sub-sample

Responsiveness of the IPAQ-SF and PASE was assessed between baseline and 3-month follow-up in a sub-group of participants, who reported they had increased their levels of PA between the baseline data collection and the 3-month follow-up on the self-reported physical activity behaviour questions. The responsiveness subsample included participants who completed either the IPAQ-SF or PASE at both baseline and 3-month follow-up and self-reported a positive change in their self-reported physical activity behaviour questions from baseline to follow-up. For example, reported not trying muscle strengthening exercises or general fitness exercises at baseline, but reported trying muscle strengthening exercises or general fitness exercises at 3-month follow-up). To allow for the analysis of responsiveness, a 3-months follow-up was selected as the most suitable second measurement time-period to evaluate whether the IPAQ-SF or PASE could detect changes in PA behaviours that occur over a sufficient intervention duration with a component that targets physical activity.

Statistical analysis

The IPAQ-SF total METS$^{-1mins-1week}$ scores were positively skewed and so a logarithmic transformation was used to allow for a parametric statistical model when evaluating measurement properties of the IPAQ-SF. Baseline descriptive statistics were reported in all participants who completed a baseline IPAQ-SF or PASE and also in the reliability and responsiveness sub-samples using frequencies with proportions or mean values with standard deviation (SD). Median and interquartile ranges (IQR) were reported for skewed data. Changes in longitudinal descriptive statistics were reported as mean change from baseline to 3-month follow-up in the reliability and responsiveness sub-samples. Changes in scores from baseline to 3-month follow-up in the reliability and responsiveness sub-samples were tested using paired sample t-tests with an α level of 5%. All analyses were conducted using Stata Statistical Software: Release 13 (StataCorp. 2013. College Station, TX: StataCorp LP).

Comparison of IPAQ-SF and PASE

Baseline data of all participants were used to compare levels of PA determined by the IPAQ-SF and PASE. Total scores of both measures, using IPAQ-SF logarithmic transformation, were compared using Pearson's correlations. Sub-domains of PA (sitting, walking, moderate intensity and vigorous intensity activities) were compared using a Spearman's rank coefficient. A priori hypothesis were made, as recommended, on the comparisons of IPAQ-SF and PASE scores (Terwee et al., 2007). We hypothesised that the IPAQ-SF and PASE would correlate in terms of total PA score and sub-domains of PA with a strong association (0.6-0.9) based on recommendations that correlations between self-report measures of PA have previously been shown to be strong (Svege, Kolle, & Risberg, 2012; Terwee et al., 2010).

Reliability and measurement error

To assess reliability of both measures, a two-way random effects intraclass correlation for absolute agreement was used ($ICC_{agreement}$)(Shrout & Fleiss, 1979). An a priori cut-off of 0.7 was selected to represent adequate reliability for both the IPAQ-SF and PASE (Terwee et al., 2007). Reliability of the IPAQ-SF categorical scoring was assessed using a quadratic weighted Kappa (Cohen, 1968). Weightings were assigned as follows; 0 for same category, 1 for adjacent categories, and 4 for 2 categories apart. Reliability was assessed in sub-domains of the IPAQ-SF and PASE. For the IPAQ-SF, vigorous, moderate, light, walking and sitting activities were tested. For the PASE, strenuous, moderate, light, walking, sitting and strengthening exercises were tested. For dichotomous items of the PASE's household and work-related activities, reliability was tested using Kappa tests and 95% confidence intervals (95%CI). Interpretation of Kappa values followed recommended reference values (Landis & Koch, 1977).

Measurement error was assessed using standard error measurement (SEM), smallest detectable change (SDC) and Bland and Altman plots. The SEM was calculated for total scores of both measures using SEM absolute agreement, to account for systematic difference between time points and residual variance. The SDC was calculated using the SEM absolute agreement. Bland and Altman plots were used to display measurement error using mean scores at baseline and 3-month follow-up and difference in scores between baseline and 3-month follow-up. For the Bland and Altman plot, 95%CI of mean values were calculated to represent 95% limits of agreement (Bland & Altman, 1986).

Responsiveness

To assess responsiveness, change scores in the PASE and IPAQ-SF were compared using effect sizes (ES), standardised responsiveness ratio (SRM) and response ratio (RR). ES were calculated by the difference in mean change divided by the baseline SD. SRM was calculated as mean change divided by the SD of change. RR was calculated as the mean change divided by the SD of the baseline of the reliability subsample. ES and SRM were interpreted using cut-off values; small (0.2-0.5), moderate (0.5-0.8) and large (0.8 or greater) (Guyatt, Walter, & Norman, 1987; Kazis, Anderson, & Meenan, 1989; Liang, Fossel, & Larson, 1990). A RR of above 1 indicated good responsiveness.

Ethical approval

The MOSAICS study was approved by the North West 1 Research Ethics Committee, Cheshire (REC reference: 10/H1017/76).

Results

Of the 525 participants who returned the MOSAICS baseline questionnaire; 489 (93%) completed either the IPAQ-SF (n=371, 70%) or the PASE (n=432, 82%) at baseline and were included in this secondary data analysis. 314 (60%) completed both the IPAQ-SF and the PASE questionnaire (60% of the total sample). Of the 470 participants who returned the MOSAICS 3-months follow-up questionnaire, there were 401 participants classified into the reliability sub-sample, who were deemed to have been stable during the study period, and had completed either the PASE or IPAQ-SF. Of these, 360 (90%) completed the PASE and 312 (78%) completed the IPAQ-SF at baseline and 3-month follow-up. There were 90 participants classified into the responsiveness sub-sample, 66 of those completed the IPAQ-SF (73%) and 83 PASE (86%) at both baseline and 3 months follow-up (**figure 1**). Baseline descriptive statistics for the whole sample and each subsample are displayed in **table 1**. We compared baseline descriptive statistics of those who completed either the IPAQ-SF or PASE at baseline to those that did not complete either PA measure to assess any differences. Those who did not complete either the IPAQ-SF or PASE were older in age, had poorer mental health (as indicated by a lower MCS score in the SF-12), but were not significantly different in terms of gender distributions, pain intensity, PCS and QoL.

Comparison of IPAQ-SF and PASE

Total scores of the IPAQ-SF and PASE were moderately associated with each other (r=0.56, *p*=<0.001). Comparisons of the sub-domains showed moderate strength associations in sitting activities (r_s=0.46, *p*=<0.001), walking activities (r_s=0.57, *p*=<0.001), moderate intensity activities (r_s=0.34, *p*=<0.001) and vigorous/strenuous activities (r_s=0.39, *p*=<0.001). While all associations were statistically significant, neither the total scores nor subdomains of the IPAQ-SF and PASE demonstrated a correlation coefficient above 0.6.

Reliability and measurement error

Changes in longitudinal descriptive statistics for the reliability sub-sample are displayed in **table 2**. Within the reliability sub-sample, there were statistically significant changes in joint pain intensity, PCS and QoL. Although changes in these outcomes were not of a magnitude commonly considered to be of a minimal clinically important change (Jenkinson & Layte, 1997; Salaffi, Stancati, Silvestri, Ciapetti, & Grassi, 2004; Walters & Brazier, 2005). The mean PASE scores also changed significantly at 3-month follow-up compared to baseline in the reliability sub-sample. The intraclass correlation between baseline and 3-month follow-up for the total score of the PASE was below the 0.7 cut-off value; $ICC_{agreement}$=0.68 (0.61-0.73 95%CI, p=<0.001), SEM was 46.7 and SDC was 129.6. **Figure 2** displays the Bland and Altman plot with the lower 95% limit of agreement -117 and upper 95% limit of agreement 136, representing large measurement error and limits of agreement when considering the score range of the PASE (0-400). The intraclass correlation between baseline and 3-month follow-up for the total score of the IPAQ-SF was below the 0.7 cut-off value; $ICC_{agreement}$=0.62 (0.55-0.71 95%CI, p=<0.001), SEM was 3532.2 $METS^{-1mins-1week}$ and SDC was 9790.8 $METS^{-1mins-1week}$. **Figure 3** displays the Bland and Altman plot with the lower 95% limit of agreement -5222 $METS^{-1mins-1week}$ and upper limits of agreement $5597METS^{-1mins-1week}$, representing large measurement error and limits of agreement, considering $3000METS^{-1mins-1week}$ equates to 6-8 hours of running in a week. A quadratic weighted Kappa showed agreement between baseline and 3-month follow-up was below 0.7 cut-off: K=0.56 (0.43-0.67 95%CI). **Table 3** displays the reliability of the sub-domains within the IPAQ-SF and PASE, Spearman's rank coefficients and Kappa values ranged from 0.33-0.74 across domains in the IPAQ-SF and PASE.

Responsiveness

Changes in longitudinal descriptive statistics for the responsiveness sub-sample are displayed in **table 2**. There were no statistically significant changes in the descriptive statistics between baseline and 3-month follow-up in the responsiveness subsample. Mean changes in PASE and median change in IPAQ-SF suggested the responsiveness sub-group reduced their levels of PA between baseline and 3 months follow-up. Low scores in indicators of responsiveness were observed; the ES was -0.14 and -0.16, SRM was -0.21 and -0.21 and RR was 0.12 and 0.09 for the IPAQ-SF and PASE respectively.

Discussion

Main findings

This study has investigated the measurement properties of the IPAQ-SF and PASE in those aged 45 and over, consulting primary care with OA of the hand, hip, knee or foot. We assessed reliability, measurement error, responsiveness and compared scoring for the IPAQ-SF and PASE, assessing the ability of both measures in detecting changes in physical activity levels. When comparing total scores of the IPAQ-SF and PASE we found that both instruments correlated moderately with each other, suggesting the IPAQ-SF and PASE are moderately similar in terms of measuring total PA levels in our sample, however, the correlation strength was below the a priori cut-off of 0.6. When exploring subdomains of PA, walking activities also had a moderate correlation to each other, but sitting, moderate and vigorous activities all had weaker correlations in comparison to those found for the total scores. The PASE contains several sub-domains that the IPAQ-SF does not, differences in categorisation of activities between the IPAQ-SF and PASE may explain the higher magnitude of association in total scores but not in matched sub-domains.

While there is no complete consensus on appropriate cut-off values for ICC to demonstrate good reliability of a measurement instrument, an ICC=0.7 has been generally recommended (Terwee et al., 2007). Neither the IPAQ-SF nor PASE achieved this in 3-month test-retest assessment, although the PASE was closest. For measurement error, the IPAQ-SF statistics represent a relatively large SEM, SDC and limits of agreement, meaning an extremely large change in weekly PA levels would be required to be detected by the IPAQ-SF outside of the measurement error. The SEM, SDC and limits of agreement findings were relatively large in relation to the PASE total scale range (0-400) suggesting large measurement error. This suggests large measurement error in proportion to possible range of total scores. In IPAQ-SF and PASE responsiveness findings suggest low responsiveness. However, it is difficult to ascertain whether this was due to a true lack of change in the responsiveness subsample or due to the instruments' inability to detect change.

Comparison with other studies

Our findings comparing the PASE and IPAQ-SF total scores showing moderate correlations were lower than previously shown strong correlations (Svege et al., 2012), although previous studies showed low correlations to activity monitors (Casartelli et al., 2015; Svege et al., 2012). The findings in measurement error of the IPAQ-SF and PASE are comparable with three previous investigations (Svege et al., 2012) and PASE (Bolszak, Casartelli, Impellizzeri, & Maffiuletti, 2014; Casartelli, Bolszak, Impellizzeri, & Maffiuletti, 2015) in adults with OA, which identified that neither the IPAQ-SF nor PASE had adequate test-retest reliability (greater than or equal to 0.7) (Terwee et al., 2007). A study using the Dutch version of the IPAQ-SF in a sample after joint replacement demonstrated test-retest reliability below 0.7 (Blikman, Stevens, Bulstra, van den Akker-Scheek, & Reininga, 2013). However, there is inconsistency in the literature with some previous studies reporting adequate test-retest reliability in both measures which could potentially be explained by a shorter time gap between testing (7-10 days) or a more active approach to keeping participants stable in their levels of PA (Bolszak et al., 2014; Naal, Impellizzeri, & Leunig, 2009; Svege et al., 2012). The findings in this study on measurement error were similar to other studies on the IPAQ-SF (Blikman et al., 2013; Bolszak et al., 2014) and PASE (Casartelli et al., 2015), suggesting that comparisons of levels of PA between individuals or time-points are at risk of large measurement error when using the IPAQ-SF and PASE. The general findings on the IPAQ-SF and PASE in terms of reliability and measurement error in our study are in line with a previous systematic review on all self-report instruments measuring PA in adults with OA, showing some evidence for acceptable reliability, but large measurement error (Smith et al., 2019).

Strengths and weakness of study

A strength of this study was the large sample size and investigation of the two most commonly used self-report PA instruments in OA research (Smith et al., 2019). Sample sizes above n=100 are generally considered adequate for statistical precision in evaluating test-retest reliability and measurement error and comparing scores of the IPAQ-SF and PASE (Terwee et al., 2007). The characteristics of primary care practices included within the MOSAICS cluster RCT are similar to those within the wider UK, allowing for our findings to be generalisable to other UK general practices (K. Dziedzic et al., 2018). In those that did not complete either the PASE or IPAQ-SF, we found only small differences in descriptive characteristics to those that completed either instrument, suggesting that the findings in our sample are generalisable to those aged 45 and over with OA in the wider primary care MOSAICS sample.

Despite the strengths of this study, the evaluation of test-retest reliability and responsiveness does have limitations. For example, our criteria for stability in the reliability subsample could not guarantee that the whole sub-sample remained stable with regards to levels of PA during the study period. This may have led to an underestimation on the instruments' reliability and measurement, if participants who were not stable in their level of PA were included into the subsample. Conversely, our criteria for changes in PA behaviours in our responsiveness subsample may not have been sensitive or specific enough to identify those that experienced true changes in their PA levels during the study period. There is a risk of underestimation of the instruments' responsiveness if those with true changes to their PA levels were not detected. An anchor measurement, such as an objective measure in levels of PA would have been desirable to accurately estimate stability or changes in true levels of PA in the study sample. However objective measures such as activity monitors are costly for population level studies and so were not viable within the MOSAICS study. The 'Gold Standard' measure of PA, doubly-labelled water (Schuit, Schouten, Westerterp, & Saris, 1997), is costly and requires specialised laboratory expertise and equipment, which were not viable as part of the MOSAICS study. Because of this, we were unable to establish the degree to which the IPAQ-SF or PASE represent true levels of PA.

Implications for research and clinical practice

Currently, there are no international recommendations to guide the selection of instruments to measure PA in adults with OA. Due to the large measurement error of the PASE and IPAQ-SF, relative to their scoring range or quantity of PA measured in METS respectively, both instruments may perform poorly when comparing individual or population PA levels or evaluating change in PA levels over time. For example, biased associations or inferences may occur in research studies that investigate the associations between PA level and other clinical outcomes or studies that investigate change in PA over time.

More evaluations of measurement properties of self-report physical activity instruments are warranted, particularly those that assess criterion validity in adults with OA. Further evaluations on test-retest reliability and responsiveness using objective measures as an anchor for detecting stability or changes in PA would also provide important information on their measurement properties. Such evaluations are necessary before recommendations can be made on the selection of instrument for measuring levels of PA in adults with OA.

Conclusion

Despite their low cost and ease to administer in population level research, the IPAQ-SF and PASE appear to be limited in their reliability and measurement error for measuring levels of PA among adults with OA. Their ability to accurately compare PA levels between populations and to detect changes in PA over time is questionable. Researchers and clinicians should be aware of the limitations of the IPAQ-SF and PASE's measurement properties for assessing PA levels among adults aged 45 and over with OA.

References

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307-310.

Blikman, T., Stevens, M., Bulstra, S. K., van den Akker-Scheek, I., & Reininga, I. H. (2013). Reliability and validity of the Dutch version of the International Physical Activity Questionnaire in patients after total hip arthroplasty or total knee arthroplasty. *journal of orthopaedic & sports physical therapy, 43*(9), 650-659.

Bolszak, S., Casartelli, N. C., Impellizzeri, F. M., & Maffiuletti, N. A. (2014). Validity and reproducibility of the Physical Activity Scale for the Elderly (PASE) questionnaire for the measurement of the physical activity level in patients after total knee arthroplasty. *BMC musculoskeletal disorders, 15*(1), 1.

Booth, M. L., Ainsworth, B. E., Pratt, M., Ekelund, U., Yngve, A., Sallis, J. F., & Oja, P. (2003). International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc, 195*(9131/03), 3508-1381.

Casartelli, N. C., Bolszak, S., Impellizzeri, F. M., & Maffiuletti, N. A. (2015). Reproducibility and Validity of the Physical Activity Scale for the Elderly (PASE) Questionnaire in Patients After Total Hip Arthroplasty. *Physical therapy, 95*(1), 86-94. doi:10.2522/ptj.20130557

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213.

Dziedzic, K., Healey, E., Porcheret, M., Afolabi, E., Lewis, M., Morden, A., . . . Finney, A. (2018). Implementing core NICE guidelines for osteoarthritis in primary care with a model consultation (MOSAICS): a cluster randomised controlled trial. *Osteoarthritis and Cartilage, 26*(1), 43-53.

Dziedzic, K. S., Healey, E. L., Porcheret, M., Ong, B. N., Main, C. J., Jordan, K. P., . . . Morden, A. (2014). Implementing the NICE osteoarthritis guidelines: a mixed methods study and cluster randomised trial of a model osteoarthritis consultation in primary care-the Management of OsteoArthritis In Consultations (MOSAICS) study protocol. *Implementation Science, 9*(1), 95.

EuroQol Group. (1990). EuroQol--a new facility for the measurement of health-related quality of life. *Health policy (Amsterdam, Netherlands), 16*(3), 199.

Felson, D. T. (2009). Developments in the clinical understanding of osteoarthritis. *Arthritis research & therapy, 11*(1), 203.

Finney, A., Dziedzic, K. S., Lewis, M., & Healey, E. (2017). Multisite peripheral joint pain: a cross-sectional study of prevalence and impact on general health, quality of life, pain intensity and consultation behaviour. *BMC musculoskeletal disorders, 18*(1), 535.

Fransen, M., McConnell, S., Harmer, A. R., Van der Esch, M., Simic, M., & Bennell, K. L. (2015). Exercise for osteoarthritis of the knee: a Cochrane systematic review. *Br J Sports Med, 49*(24), 1554-1557.

Fransen, M., McConnell, S., Hernandez-Molina, G., & Reichenbach, S. (2014). Exercise for osteoarthritis of the hip. *Cochrane Database of Systematic Reviews*(4).

Grime, J., & Dudley, B. (2014). Developing written information on osteoarthritis for patients: facilitating user involvement by exposure to qualitative research. Health Expectations, 17(2), 164-173.

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of chronic diseases, 40*(2), 171-178.

Healey, E. L., Allen, K. D., Bennell, K., Bowden, J. L., Quicke, J. G., & Smith, R. (2020). Self-Report Measures of Physical Activity. *Arthritis Care & Research, 72*, 717-730.

Herbolsheimer, F., Schaap, L. A., Edwards, M. H., Maggi, S., Otero, Á., Timmermans, E. J., . . . Cooper, C. (2016). Physical Activity Patterns Among Older Adults With and Without Knee Osteoarthritis in Six European Countries. *Arthritis Care & Research, 68*(2), 228-236.

Holden, M. A., Nicholls, E. E., Young, J., Hay, E. M., & Foster, N. E. (2015). Exercise and physical activity in older adults with knee pain: a mixed methods study. Rheumatology, 54(3), 413-423.

Jenkinson, C., & Layte, R. (1997). Development and testing of the UK SF-12. *Journal of health services research & policy, 2*(1), 14-18.

Jinks, C., Jordan, K., Ong, B., & Croft, P. (2003). A brief screening tool for knee pain in primary care (KNEST). 2. Results from a survey in the general population aged 50 and over. *Rheumatology, 43*(1), 55-61.

Juhl, C., Christensen, R., Roos, E. M., Zhang, W., & Lund, H. (2014). Impact of exercise type and dose on pain and disability in knee osteoarthritis: a systematic review and meta-regression analysis of randomized controlled trials. *Arthritis & Rheumatology, 66*(3), 622-636.

Kazis, L. E., Anderson, J. J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. *Medical care*, S178-S189.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Liang, M. H., Fossel, A. H., & Larson, M. G. (1990). Comparisons of five health status instruments for orthopedic evaluation. *Medical care*, 632-642.

Naal, F. D., Impellizzeri, F. M., & Leunig, M. (2009). Which is the best activity rating scale for patients undergoing total joint arthroplasty? *Clinical orthopaedics and related research, 467*(4), 958-965.

National Institute for Health and Care Excellence (2014). Osteoarthritis: care and management.

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity, 5*(1), 1.

Quicke, J. G., Foster, N. E., Croft, P. R., Ogollah, R. O., & Holden, M. A. (2018). Change in physical activity level and clinical outcomes in older adults with knee pain: a secondary analysis from a randomised controlled trial. *BMC musculoskeletal disorders, 19*(1), 59.

Salaffi, F., Stancati, A., Silvestri, C. A., Ciapetti, A., & Grassi, W. (2004). Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *European Journal of Pain, 8*(4), 283-291.

Schuit, A. J., Schouten, E. G., Westerterp, K. R., & Saris, W. H. (1997). Validity of the Physical Activity Scale for the Elderly (PASE): according to energy expenditure assessed by the doubly labeled water method. *Journal of clinical epidemiology, 50*(5), 541-546.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin, 86*(2), 420.

Silsbury, Z., Goldsmith, R., & Rushton, A. (2015). Systematic review of the measurement properties of self-report physical activity questionnaires in healthy adult populations. *BMJ open, 5*(9), e008430.

Smith, R. D., Dziedzic, K. S., Quicke, J. G., Holden, M. A., McHugh, G. A., & Healey, E. L. (2019). Identification and Evaluation of Self-Report Physical Activity Instruments in Adults With Osteoarthritis: A Systematic Review. *Arthritis Care & Research, 71*(2), 237-251.

Svege, I., Kolle, E., & Risberg, M. A. (2012). Reliability and validity of the Physical Activity Scale for the Elderly (PASE) in patients with hip osteoarthritis. *BMC musculoskeletal disorders, 13*(1), 1.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., . . . de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology, 60*(1), 34-42.

Terwee, C. B., Mokkink, L. B., van Poppel, M. N., Chinapaw, M. J., van Mechelen, W., & de Vet, H. C. (2010). Qualitative attributes and measurement properties of physical activity questionnaires. *Sports Medicine, 40*(7), 525-537.

Uthman, O. A., van der Windt, D. A., Jordan, J. L., Dziedzic, K. S., Healey, E. L., Peat, G. M., & Foster, N. E. (2013). Exercise for lower limb osteoarthritis: systematic review incorporating trial sequential analysis and network meta-analysis. *Bmj, 347*, f5555.

Walters, S. J., & Brazier, J. E. (2005). Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life Research, 14*(6), 1523-1532.

Warburton, D. E., Nicol, C. W., & Bredin, S. S. (2006). Health benefits of physical activity: the evidence. CMAJ, 174(6), 801-809.

Ware Jr, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Medical care, 34*(3), 220-233.

Washburn, R. A., Smith, K. W., Jette, A. M., & Janney, C. A. (1993). The Physical Activity Scale for the Elderly (PASE): development and evaluation. *Journal of clinical epidemiology, 46*(2), 153-162.

Table 1 Baseline descriptive statistics of whole sample, those that completed the IPAQ-SF and PASE at baseline, reliability subsample and responsiveness subsample.

	Baseline IPAQ-SF and PASE (n=314)	**Reliability subsample (n=401)**	**Responsiveness subsample (n=90)**
Gender, n (%)			
Females	172 (55)	530 (57)	50 (56)
Site of peripheral joint pain, n (%)			
Knee	259 (82)	335 (84)	71 (79)
Hip	181 (57)	232 (58)	46 (51)
Feet	150 (48)	194 (48)	42 (47)
Hand	167 (53)	220 (55)	54 (60)
Mean age, years (SD)	67.2 (10.5)	68.3 (10.6)	66.8 (10.0)
Mean BMI, kg/m^2 (SD)	27.0 (8.1)	28.4 (4.9)	29.7 (6.2)
Mean pain intensity (SD)	7.4 (2.1)	7.4 (2.0)	6.7 (2.3)
Mean health Status, SF-12 (SD)			
PCS	36.7 (11.3)	36.8 (11.5)	38.1 (9.2)
MCS	51.5 (11.3)	51.2 (10.9)	51.8 (11.3)
Mean QoL, EQ-5D (SD)	0.6 (0.3)	0.6 (0.3)	0.6 (0.3)
Mean PASE score (SD)	142.3 (78.8)	140.3 (76.3)	153.4 (89.3)
Median IPAQ-SF, total METS$^{-1mins-1week}$ (interquartile range)	1527 (462-3732)	1386 (198-3452)	2574 (305-5153)
IPAQ-SF categories, n (%)			
Low	113 (36)	112 (36)	17 (26)
Moderate	103 (33)	104 (33)	21 (32)
High	98 (31)	96 (31)	28 (42)

Note: percentages may not equal 100% due to rounding up. Abbreviations: SD, standard deviation, IPAQ-SF, International physical activity questionnaire - short form, PASE, physical activity scale for the elderly, BMI, body mass index, SF-12, 12-Item Short Form Survey, PCS, physical component score, MCS, mental component score, QoL, quality of life, METS$^{-1mins-1week}$, metabolic equivalent per minute per week.

Table 2 Changes in longitudinal descriptive statistics from baseline to 3 months follow-up for reliability and responsiveness subsample.

	Change in reliability subsample (n=401)	***p*-value**	**Change in responsiveness subsample (n=90)**	***p*-value**
Change in mean BMI, kg/m^2	-0.1	0.19	-0.5	0.18
Change in mean pain intensity	-1.3	<0.001*	-0.5	0.06
Change in mean health Status, SF-12				
PCS	1.0	0.03*	-0.4	0.7
MCS	0.4	0.81	-0.6	0.6
Change in mean QoL, EQ-5D	0.1	<0.001*	-0.1	0.3
Change in mean PASE score	-8.8	0.04*	-14.3	0.2
Change in mean IPAQ-SF, total METS$^{-1mins-1week}$	323	0.52	-347	0.1
Global assessment of change, n (%)				
Missing	9 (2)		0 (0)	
Completely recovered	5 (1)		1 (1)	
Much better	40 (10)		10 (11)	
Better	80 (20)		20 (22)	
No change	141 (35)		39 (43)	
Worse	108 (27)		16 (18)	
Much worse	18 (4)		4 (4)	

Note: percentages may not equal 100% due to rounding up. Abbreviations: IPAQ-SF, International physical activity questionnaire - short form, PASE, physical activity scale for the elderly, BMI, body mass index, SF-12, 12-Item Short Form Survey, PCS, physical component score, MCS, mental component score, QoL, quality of life, METS$^{-1mins-1week}$, metabolic equivalent per minute per week.

Table 3 Reliability of PASE and IPAQ-SF subdomains in the reliability subsample.

Instrument subdomain	**Coefficient**
PASE – Leisure time activities	**Spearman's Rank, *p* value**
Sitting activities	r=0.54, *p*=<0.001
Walking activities	r=0.57, *p*=<0.001
Light intensity activities	r=0.33, *p*=<0.001
Moderate intensity activities	r=0.47, *p*=<0.001
Strenuous intensity activities	r=0.64, *p*=<0.001
Muscle endurance or strengthening activities	r=0.56, *p*=<0.001
PASE - Household and work-related activities	**Weighted Kappa (95%CI)**
Light housework	K=0.49 (0.08-0.89)
Heavy housework	K=0.60 (0.48-0.72)
Home repairs	K=0.49 (0.32-0.65)
Garden care	K=0.46 (0.34-0.58)
Outdoor gardening	K=0.37 (0.25-0.50)
Caring others	K=0.52 (0.39-0.64)
Work or volunteer	K=0.65 (0.53-0.77)
Working physical activity	K=0.50 (0.29-0.71)
IPAQ-SF – Subdomains	**Spearman's Rank, *p* value**
Sitting activities	r=0.74, *p*=<0.001
Walking activities	r=0.59, *p*=<0.001
Moderate intensity activities	r=0.45, *p*=<0.001
Vigorous intensity activities	r=0.46, *p*=<0.001

Abbreviations: IPAQ-SF, International physical activity questionnaire - short form, PASE, physical activity scale for the elderly.

Figure 1 Flowchart of study participants for analysis taken from the MOSAICS study.

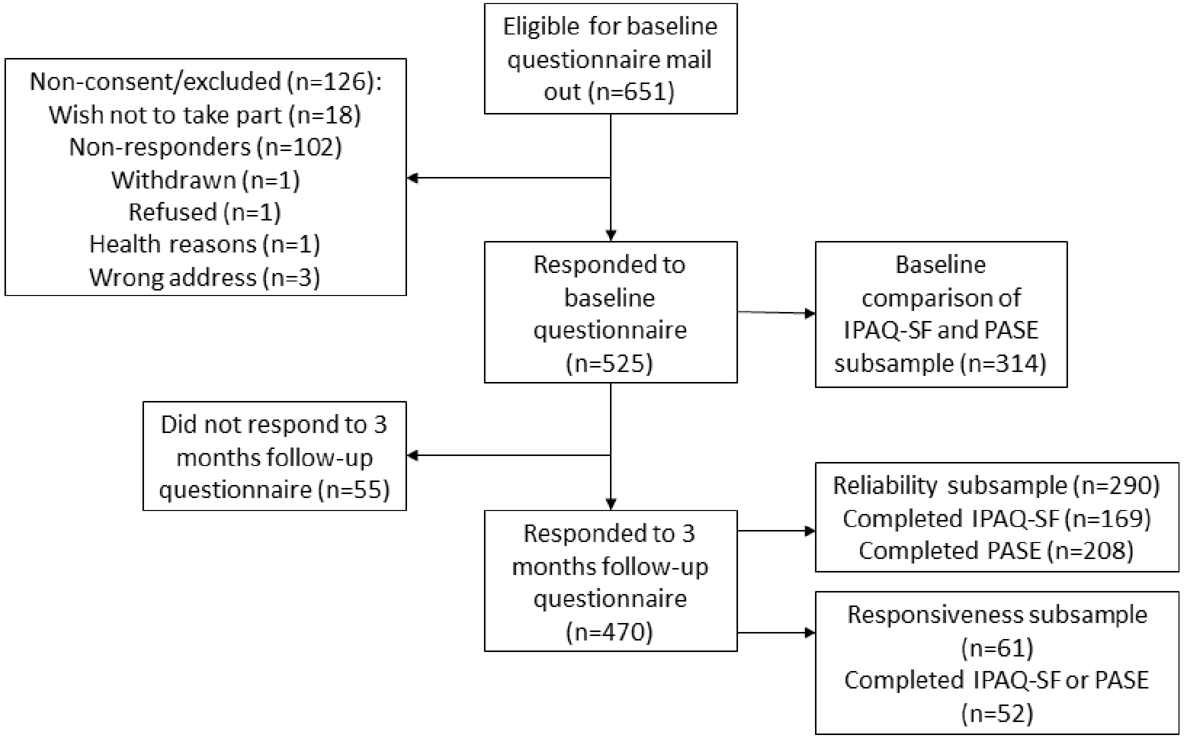

Eligible for baseline questionnaire mail out (n=651)
Non-consent/excluded (n=126):
Wish not to take part (n=18)
Non-responders (n=102)
Withdrawn (n=1)
Refused (n=1)
Health reasons (n=1)
Wrong address (n=3)
Responded to baseline questionnaire (n=525)
Baseline comparison of IPAQ-SF and PASE subsample (n=314)
Did not respond to 3 months follow-up questionnaire (n=55)
Responded to 3 months follow-up questionnaire (n=470)
Reliability subsample (n=290)
Completed IPAQ-SF (n=169)
Completed PASE (n=208)
Responsiveness subsample (n=61)
Completed IPAQ-SF or PASE (n=52)

Figure 2 Bland and Altman plot of the PASE in the reliability subsample from baseline to 3-months follow-up.

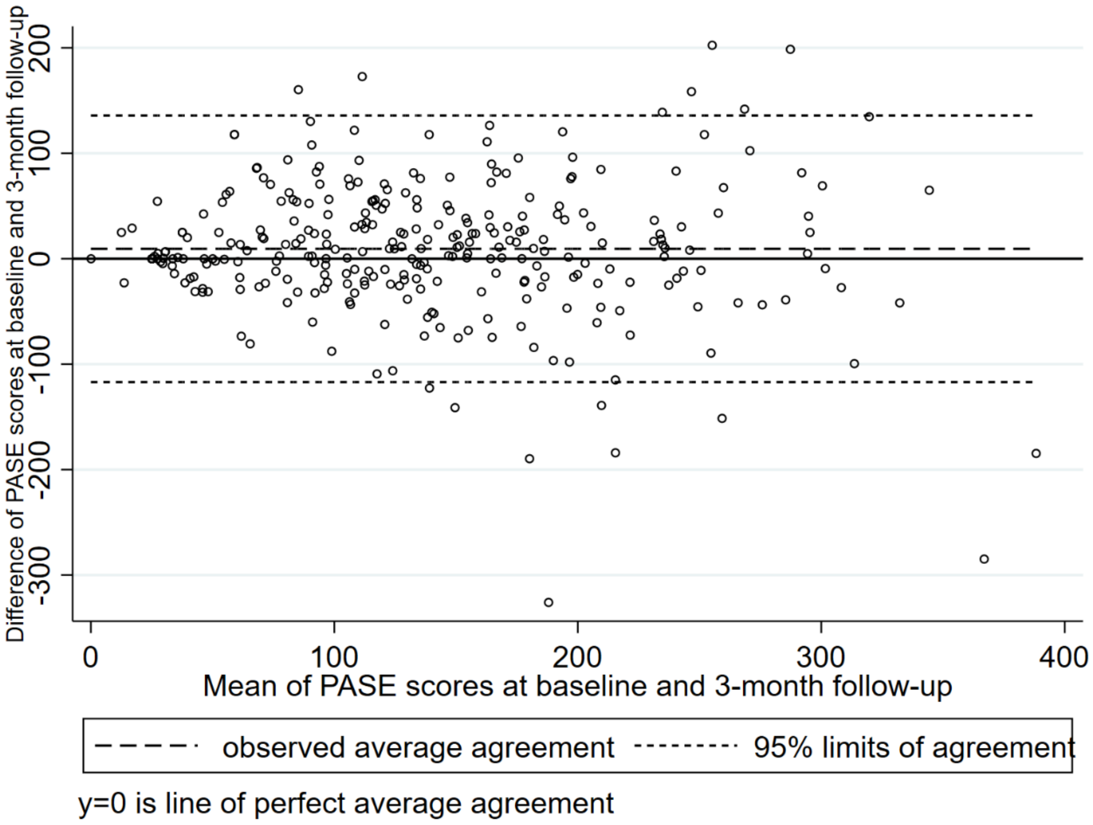
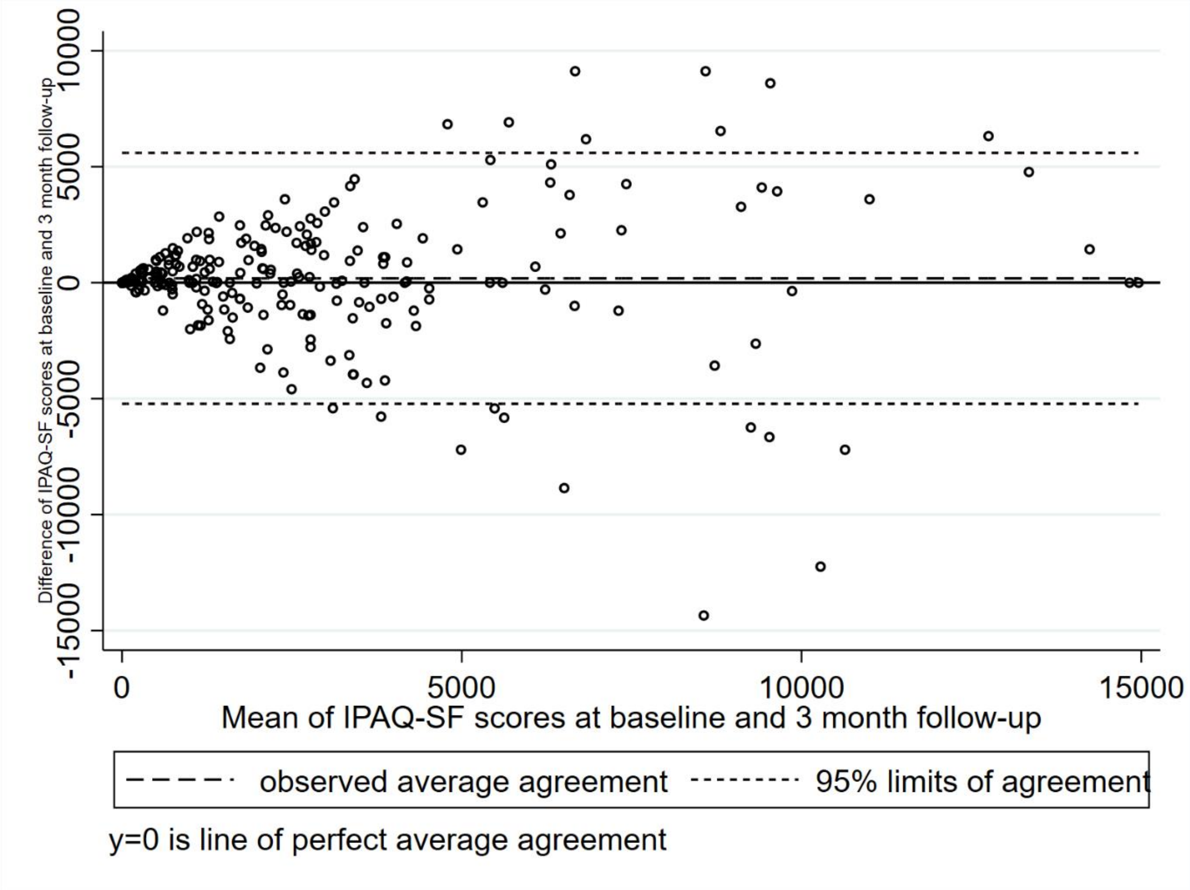


Figure 3 Bland and Altman plot of the IPAQ-SF in the reliability subsample from baseline to 3-months follow-up.


Difference of IPAQ-SF scores at baseline and 3 month follow-up
Mean of IPAQ-SF scores at baseline and 3 month follow-up
observed average agreement
95% limits of agreement
y=0 is line of perfect average agreement