Translation of the International Physical Activity Questionnaire to Maltese and reliability testing

Spiteri, K., Grafton, K., Xerri de Caro, J. & Broom, D.

Original citation & hyperlink:
Spiteri, K, Grafton, K, Xerri de Caro, J & Broom, D 2021, 'Translation of the International Physical Activity Questionnaire to Maltese and reliability testing', Journal for the Measurement of Physical Behavior, vol. 4, no. 1.
https://dx.doi.org/ 10.1123/jmpb.2020-0031

DOI 10.1123/jmpb.2020-0031
ISSN 2575-6605
ESSN 2575-6613

Publisher: Human Kinetics

Translation of the International Physical Activity Questionnaire to Maltese and reliability testing

Abstract

The International Physical Activity Questionnaire (IPAQ) is a widely used self-reported physical activity (PA) measure at population level, developed to allow for international cross-country comparisons. Due to its unavailability, the aim of this study was to translate the IPAQ-long to Maltese and undertake reliability testing. The IPAQ-long English version was translated into Maltese following the IPAQ guidelines which included backwards translation. Maltese speaking participants, aged between 18 and 69 years, were recruited through convenience sampling (n = 170). Participants completed the IPAQ-long twice within an 8 to 48 hour period between completions. PA was calculated in MET minutes per week and reliability was calculated using Spearman correlation, interclass correlation coefficient (ICC), concordance correlation coefficient (CCC) and Bland Altman plots. 155 participants completed the questionnaire at two time points. Spearman correlation was 0.83 (0.76-0.88) for total PA and 0.84 (0.77-0.89) for total sitting time. The ICC was 0.83 (0.76-0.88) and CCC 0.75-0.87 for total PA. The lowest reliability was for total transport with CCC of 0.21-0.45. Bland Altman plots highlight that 95% of the differences fell within 2 standard deviations from the mean. Since the Maltese IPAQ-long has similar reliability to the English version, we recommend healthcare professionals and physical activity practitioners use this tool when examining population level PA amongst Maltese speaking individuals.

Introduction

The beneficial health effects of moderate-to-vigorous intensity physical activity (PA) in reducing the risk of heart disease and other chronic diseases is well known (Physical Activity Guidelines Advisory Committee, 2018). Regular PA, as a lifestyle choice, has a positive influence on the health of individuals (Lee et al., 2012). There is evidence to show that physical inactivity is linked with non-communicable disease risk progression, including, type 2 diabetes, stroke, coronary heart disease and osteoporosis (UK Chief Medical Officers, 2019). Being physically active has beneficial effects in all age groups. In children, PA improves cognitive function, bone health and weight status. During adulthood, it lowers the risk of mortality, cardiovascular related adverse events, decreases risk for certain cancers, improves cognitive function and reduces the risk of dementia. In older adults it reduces the risk of falls and improves functioning (Physical Activity Guidelines Advisory Committee, 2018).

The term PA is used to describe various aspects of daily behaviour and activities (Troiano, McClain, Brychta, & Chen, 2014). Given the complexity of PA behaviour, there are various metrics which demonstrate a beneficial health effect: a) bouts of at least 10 minutes, b) light-intensity physical activity and c) short bouts less than 10 minutes of moderate-to-vigorous PA (Physical Activity Guidelines Advisory Committee, 2018). Measurement of PA can be undertaken using device-based or self-reported measures (Troiano, Pettee Gabriel, Welk, Owen, & Sternfeld, 2012). Device-based measures include pedometers and accelerometers which are deemed to be more valid as they measure PA behaviour directly (Loney, Standage, Thompson, Sebire, & Cumming, 2011). Troiano et al., (2012) argue that device-based and self-report tools measure different aspects of PA. Whilst device-based methods measure precise quantifiable movement of a body part, self-reported methods incorporate the individual's perceptions about their PA behaviour. Even though, device-based methods are becoming more affordable and are being used in large scale cohort studies, as well for surveillance data they still remain more expensive than questionnaires (Troiano et al., 2014). Self-reported methods are more easily accessible and can be administered at a lesser expense but tend to overestimate PA levels due to recall bias, difficulty in gauging intensities and possible social desirable responses (Loney et al., 2011).

Various self-reported PA questionnaires have been developed and the majority have concurrent validity within the same range at 0.30-0.46 (Helmerhorst, Brage, Warren, Besson, & Ekelund, 2012). Concurrent validity is better with vigorous intensity activities as they are typically more structured during leisure time and therefore less prone to recall bias (Pedišić, Jurakić, Rakovac, Hodak, & Dizdar, 2011).

The International Physical Activity Questionnaire (IPAQ) is an established PA questionnaire which is used widely in the literature and researched extensively across multiple languages (Silsbury, Goldsmith, & Rushton, 2015). At the time of writing (March 2020), a search in Pubmed identified 4902 papers using IPAQ. The questionnaire was developed during 1998-99 by a group of World Health Organisation (WHO) experts whose aim was to develop a tool to encourage cross country comparisons (Bauman et al., 2009). The importance of IPAQ is its use in large scale surveys like EUPASS (Rütten et al., 2003), and Eurobarometer (Sjöström, Oja, Hagströmer, Smith, & Bauman, 2006). IPAQ has two forms being the short and long version. The short version consists of seven questions and gleans data on vigorous and moderate intensity PA, walking and sitting time on weekdays. The long version consists of 27 questions and gleans the same data but in different activities i.e. work, transport, domestic, and leisure time.

Since IPAQ-long collates data from these four different domains it can generate over-reporting of activity, as the same activity can be counted in two domains (Hallal et al., 2010).

Even though IPAQ-long measures PA behaviour in different domains, it is limited to aerobic types of PA and does not include strength or balance types of activities which are found to have a beneficial health effect (Troiano et al., 2012). The measured PA is computed into MET minutes per week with preset MET values for the different intensities. This makes it difficult to compare to national PA recommendations which are in minutes per week. METs are a measure of absolute physiological intensity, whilst self-reported methods measure relative intensity, which varies by level of fitness and health status. This creates a disparity between what is being attempted to be measured and what would have been actually measured (Troiano et al., 2012). The IPAQ-long only asks about PA in bouts of at least 10 minutes which creates problems of over-reporting, as this is based on a person's perceptions and creates inaccuracies (Hallal et al., 2010). When dealing with PA at home and at work, it is difficult to gauge intensity and bouts of activity in these domains (Sebastião et al., 2012). Another issue which can cause over-reporting in the IPAQ-long is the concept of average time spent in an activity during a week (Hallal et al., 2010). Even though the IPAQ-long has various pitfalls, there is no one tool to measure PA (Dowd et al., 2018).

As PA occurs during occupation, leisure, domestic and transport (Troiano et al., 2014), one of the advantages of IPAQ-long is that it self-reports PA behaviours in these different domains (Sebastião et al., 2012). PA at work might be an important contributor to overall PA. In a population of white collar workers, PA at work contributed to about 25% of total PA when using IPAQ-long (Kwak, Hagströmer, & Sjostrom, 2012). The correlation of occupational PA within the IPAQ-long when compared to the accelerometer, was found to be moderate (0.46) (Kwak et al., 2012). Life events such as death of a spouse, change in marital status and retirement might cause a change in PA behaviour patterns (Gropper, John, Sudeck, & Thiel, 2020). The IPAQ-long is a self-reported measure which could be used to assess PA behaviour across different domains.

The IPAQ has been translated into more than 20 different languages including Turkish, Serbian, Croatian, Nigerian, Malay, and French (IPAQ group, 2019), but as yet has not been translated into the Maltese language. It is important that each localised version has its reliability tested as recalling PA behaviour is a complex cognitive process that can generate errors because of question interpretation and cultural differences (Craig et al., 2003; Pedišić et al., 2011; Troiano et al., 2012). Craig et al.(2003) measured the reliability and validity of the different forms of IPAQ in 12 countries. Concurrent validity was estimated by comparing the IPAQ-long reports with accelerometer measurements and the correlation ranged between 0.26-0.39 (Craig et al., 2003). The reliability was checked using correlation coefficient, using the 'last 7 days' (CC = 0.79) or 'usual week' (CC = 0.69), the reference period did influence, the correlation coefficient of the IPAQ. The use of 'last 7 days' is recommended in questionnaires as it provides a reference period, the correlation coefficient in reliability and validity are better when compared to device-based methods compared to those obtained when using 'usual week' (Doma, Speyer, Leicht, & Cordier, 2017). Correlation between IPAQ-long and accelerometer measurements ranges between 0.26-0.39 (Craig et al., 2003). Subsequent to Craig et al.(2003) validation work, other researchers have translated the IPAQ into their language and tested reliability of the translated versions (Kalvenas, Burlacu, & Abu-Omar, 2016; Mannocci et al., 2014; Pedišić et al., 2011).

To the authors' knowledge there are no published studies which have reported the translation and reliability of a Maltese version of IPAQ-long. The aim of this research was to translate the English version of the IPAQ-long into Maltese and subsequently undertake reliability testing.

Methods

The study was completed in two parts: a) translation, and b) reliability testing. Participation in the study was voluntary and written consent was obtained from all participants. Ethical approval was obtained from the Faculty of Health and Wellbeing Ethics Committee, Sheffield Hallam University, UK. The Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were used (Kottner et al., 2011).

For part a) the translation process was carried out following IPAQ cultural adaptation (IPAQ group, 2019). The English version was used as the translation template (IPAQ group, 2019). Two experienced professional translators were paid to translate the English IPAQ-long into Maltese separately. These translations were reviewed by [redacted] and [redacted] to check for differences and merged. Minimal differences were identified between the two translations which were grammar related. The developed questionnaire was then distributed to three experts: 1) a public health specialist working in health promotion, 2) a physiotherapist with experience in translations and 3) a Maltese linguist. All experts were bilingual. The experts identified wording which required cultural adaptation such as the removal of train[1] and the introduction of scooter[2], as a mode of transport to obtain normative equivalence (Behling & Law, 2011). The newly translated version was backwards translated by two different experienced professional translators who were paid for their services. The resulting English and Maltese versions were reviewed by [redacted] and [redacted]. Some syntax translations were not exactly the same. These were discussed with the translators and the Maltese version was readapted.

The Maltese version was then piloted using cognitive interviewing with 10 people from different educational backgrounds. Participants were purposefully identified with different

[1] In the population where reliability was tested ‘train’ is not an available mode of transport so it was removed as an example.

[2] In the population studied ‘scooter’ refers to low powered motorcycles commonly known as a ‘moped’ in other languages, and not as the English term denotes.

educational background who were able to read and write. The participants age ranged from 25 to 67 years with an average age of 45 years (SD = ±17.4). Six participants were female, whilst four were male. Four participants were educated to primary school level and three had completed secondary school level education, see table 1 for participant characteristics. Each interview lasted between 35 to 60 minutes.

During the interview [redacted] took fieldwork notes and audio recorded the interview as a reference, in case written notes were unclear. The interview was carried out in an open-ended format. Questions were read by the participants and then they were asked to verbalise their thinking. Recall of PA behaviour is a cognitive process and using a cognitive interviewing method investigates the cognitive process of how participants recall information (Grey, 2015). When questions were not eliciting the intended cognitive process, [redacted] provided participants with optional wording to assist the intended recollection. After the initial four interviews, it was noted that some participants were finding the term ‘attivita fiziika vigoruz’(vigorous activity) difficult to understand, therefore, ‘attivita iebsa’ was added. As the Maltese population typically uses English words during daily language, the terms ‘weekend’ and ‘weekday’ were included in the translation as reference. This process assisted in improving the translation. The sample size for the cognitive interviewing was not set, but continued until no further changes were identified. This was done to ensure that the questionnaire was eliciting the expected response. Since the Maltese language lacks certain semantic equivalence (Behling & Law, 2011) during the translation process changes were carried out to the Maltese version which were discussed between [redacted] and [redacted] then, confirmed with one of the translators to ensure that appropriate semantic meaning was maintained.

For part b) reliability testing; participants between 18 and 69 years were recruited. The IPAQ was developed for this target population as older adults might require different prompts to recall their PA behaviour (Bauman et al., 2009). Participants had to be a) able to read and write b) be comfortable replying to the questionnaire in the Maltese language, and, c) not have any form of disability that would limit participation in daily PA such as walking. Participants were recruited using convenience sampling. Initial sampling was from hospital workers and approaching people in the street. Snowballing was then used to reach the required quota. Recruitment was voluntary and participants could withdraw at any time.

The required sample size was calculated using Walter, Eliasziw and Donner (1998) formula based on the hypothesised intraclass correlation coefficient (ICC) values, the minimal expected ICC value and the number of observations as recommended by Streiner et al.(2015). The expected ICC for IPAQ-long based on Craig et al.(2003) was 0.8, the minimal expect level is 0.7, the number of observations are 2. This gave a sample size of 116. Previous IPAQ reliability studies had approximately 5-15% non-completion rates (MacFarlane, Chan, & Cerin, 2011; Milanovic et al., 2014; Pedišić et al., 2011). Considering a 15% non-completion rate, the expected final sample was calculated at 134 participants.

There is no fixed retest period for assessing reliability, the recall period between test and retest needs to be long enough to decrease the risk of recall bias, but short enough to measure the same behaviour (Henrica, 2011). Using a longer retest period would decrease the reliability of the questionnaire as highlighted in other studies (Helmerhorst et al., 2012; Pedišić et al., 2011). PA is a varied, multidimensional behaviour which varies from day to day, week to week and within and between seasons (Bergman, 2018; Kelly, Fitzsimons, & Baker, 2016). Certain PA behaviours might be stable over a period of a week, while others are not (Bergman, 2018; Bergman & Hagströmer, 2020). This would influence the reliability of the tool even when using devised-based measures which are considered more accurate (Bergman, 2018; Bergman & Hagströmer, 2020). To try and assess the same behavioural pattern the retest timeframe was kept to a short period, between 8 to 48 hours, and "last 7 days" was used as the recall cue to keep the same reference period. As the test retest period was kept short, there might have been a risk for the participants recalling the questions (Helmerhorst et al., 2012). To decrease the risk of recall bias, the reliability of IPAQ-long was tested together with other questions. The final questionnaire was eight pages long, included 128 questions, and it took between 20 to 40 minutes to complete. This aimed to reduce recall bias when carrying out the retest. Demographic data on age, gender, education and self-reported height and weight were collected. Height and weight were used to calculate BMI - weight (kg) / height $(m)^2$. The first questionnaire was not completed on a weekend day. This was done as PA behaviour might vary during the weekend (Nordman, Matthiessen, Biltoft-Jensen, Ritz, & Hjorth, 2020). However, the retest would have been completed on a weekend if the initial questionnaire was completed on a Thursday or Friday.

Data analysis was carried out in accordance with the IPAQ guidelines (IPAQ group, 2019). A Microsoft Excel © version 2007 spreadsheet was developed to analyse the data. IBM SPSS © version 24 was used for statistical analysis. The IPAQ-long gleans information about four different domains and activities at different intensities. A total of 11 separate activity categories were obtained. Each category was expressed in metabolic equivalent task (MET) minutes per week. The total time spent in each category was multiplied by the intensity of the activity. Walking was taken as 3.3 MET, moderate intensity PA at 4 MET and vigorous intensity PA at 8 MET. Each physical activity score was calculated and then summed to obtain amount of METs per week for walking, moderate and vigorous intensity activity. When summed these three intensities for each domain provided a total MET minutes per week score. IPAQ-long measures sitting time, it captures sitting time during transport and during weekdays and weekends. Total sitting time was calculated by multiplying time by days which gave total minutes per week. A detailed description of the IPAQ-long data processing can be found in the official guidelines (IPAQ group, 2019). Demographic characteristics were analysed descriptively. Mean and standard deviation are presented. Data cleaning was carried out following IPAQ guidelines. Twelve questionnaires were excluded due to missing data. If activity duration was lower than ten minutes these were also excluded. The primary aim of the study was not to assess the PA level within the population but the reliability of the tool, therefore data truncation was not carried out as suggested by the IPAQ guidelines. This was done in order to assess for the true difference between the test and retest. Reliability of the IPAQ- long was assessed using the ICC, standard error of the mean, standard error of measurement (SEM), Lin's concordance correlation coefficient (CCC) and Bland and Altman's plots. All variables were tested for normality using the Kolmogorov-Smirnov test. When data was not normally distributed, non-parametric tests were used. Prior to checking level of agreement, correlation between the variables was checked. Pearson correlation was used to check the correlation coefficient, which if not significant, further testing was not performed. If the correlation was significant, difference between test and retest were checked for normal distribution to check that ICC assumption were met (supplementary material 1).

The ICC takes into account the differences between the means of the measures being considered to assess the level of agreement between the two tests. CCC is used to assess the level of agreement and disagreement between the two tests. Since CCC does not assume a common mean to assess level of agreement, it can be used when the tests have different means and variances, therefore, it does not require normal distribution of variables (Liu et al., 2016). Bland and Altman's plots were used to check for repeatability of measures by plotting the mean differences between the two measures. 95% of the difference should be within 2 standard deviations of the mean difference for the tool to have good repeatability (Altman & Bland, 1986). The SEM is the standard deviation of the measurement error (Thompson & Wesolowski, 2018). It measures how far apart the outcome of repeated measure is around a single measurement, the smaller the SEM does not automatically mean better reliability (Henrica, 2011). The IPAQ-long provides the PA categories in addition to MET hours per week, so Cohen Kappa was used to check for the repeatability of the tool when categorising participants by PA levels (Sim & Wright, 2005).

Results

A total of 160 questionnaire packs were distributed and 136 participants completed the first questionnaire giving an 85% completion. The completion rate of the two questionnaires was 72% (n = 115). 17 of the questionnaires were collected online and the remaining were collected using hard-copies. Participants used the same mode of administration for both tests. The age range of participants was between 18 and 69 years with a mean of 39 years (SD = ±14). 61% of the participants were female. 68% of the participants had a tertiary level of education with the remaining had secondary education or less. 44% of the participants were married, and 50% were single. The mean Body Mass Index (BMI) was 25.3 kg/m^2 (SD = ±4.4).

There were no significant differences in BMI, education, PA and age between genders. No statistical difference in total PA was found with age ($p = 0.96$) and marital status ($p = 0.79$). Near significant difference was found in total PA with education ($p = 0.05$), with a higher mean PA in participants with lower education. Participants with lower education had significantly higher levels of MET minutes per week in transport PA ($p < 0.01$) and MET minutes per week in total walking ($p < 0.01$) and MET minutes per week in total moderate ($p < 0.01$) compared to others. Table 2 shows PA levels inter quartile ranges in different domains for test and retest. The largest difference in distribution between test and retest was for work PA, followed by leisure time PA. The median for other PA domains was similar between the test and retest. Before undertaking reliability testing the difference between t1 and t2 was checked for normal distribution, supplementary material 1 shows distribution of difference for total PA in MET minutes per week. All PA data was normally distributed and reliability testing was carried out using the ICC two way mixed model effect (Koo & Li, 2016), SEM (Thompson & Wesolowski, 2018) and CCC (Liu et al., 2016).

The ICC for all PA domains ranged between 0.7 to 0.88. The exception was total transport PA, which had poor test retest correlation of 0.34 (0.17 – 0.49). The SEM for total vigorous PA was 601 MET min per week (1.25 hours of vigorous activity per week) and for walking PA 1141 MET min per week (5.77 hours of walking per week). SEM for total sitting time was 182 minutes per week (3 hours per week). ICC of total PA was statistically significant at 0.83 (0.76 - 0.88), CCC was 0.76 – 0.88. ICC for total sitting time was 0.88 (0.83 – 0.92) and CCC was 0.83 – 0.91. Both of these variables showed good reliability for test retest. Table 3 shows the reliability statistical calculations for each of the IPAQ-long variables evaluated.

Bland-Altman plots for total PA and total sitting time (Figures 1 and 2) showed more than 95% of the variables falling within two standard deviations. The plots for all of the variables are available in the supplementary material. All variables plotted had 95% of variables falling within two standard deviations. Cohen Kappa for PA categories from IPAQ-long was 0.60 ($p < 0.01$).

Discussion

The aim of this study was to translate and assess the reliability of the translated Maltese version of IPAQ-long (MT-IPAQ-long) self-administered form. To the authors knowledge this is the first published reliability study of a Maltese translation of the IPAQ-long. Spearman's correlation coefficient for total PA and total sitting of the MT-IPAQ-long was similar to that obtained by Craig et al. (2003) in the 12-country reliability testing. This shows that the reliability is similar to the original English version. The Spearman's correlation coefficients for moderate and vigorous intensity PA were similar to the English IPAQ-long. In other studies, vigorous intensity PA had the highest reliability as this type of activity is mostly based on leisure time PA with less possibility of recall bias (Chu & Moy, 2015; Pedišić et al., 2011). In the current study, leisure time PA had the highest correlation, however, contrary to other studies domestic PA had the highest ICC and the lowest SEM. This type of PA is usually not recalled as well as leisure time PA, due to possible variability within the same week (Sebastião et al., 2012). The mean and median domestic PA levels were higher in the current population compared to other studies (Chu & Moy, 2015; Oyeyemi et al., 2014). Reliability is population specific (Kelly et al., 2016), variations in the type of activity undertaken by the population understudy might be another reason for these differences.

Spearman's correlation coefficient is not a good measure of reliability when interpreted in isolation as it does not take into consideration rater bias which is accounted for in the calculation of ICC (Bruton, Conway, & Holgate, 2000; Liu et al., 2016). Therefore ICC, CCC and Bland-Altman plots were used to interpret the reliability of the tool.

The MT-IPAQ-long highest ICC values were obtained for total domestic (0.83), with the lowest values obtained for transport (0.34). All obtained values for MT-IPAQ-long, except for transport PA had an ICC above 0.7 which shows that the tool is reliable. Using MT-IPAQ-long, worse reliability was obtained for transport when compared to other studies (Chu & Moy, 2015; MacFarlane et al., 2011; Oyeyemi et al., 2014). This study used a short recall period of up to 48 hours when the retest took place. It would be expected that the short retest period would result in higher reliability measures due to question recall. The inclusion of other questions to decrease recall bias could have been effective in reducing question recall. Reliability could have been influenced by the low PA in transport with the median being of one hour per week of PA in transport.

Comparing to other country translation studies, the ICC of MT-IPAQ-long obtained higher ICC (0.85) for total PA compared to the Arabic version (ICC = 0.66) but lower for vigorous activity (MT ICC= 0.84, Arabic ICC = 0.96) (Helou et al., 2017). In the latter study, the recall period was three weeks, the average MET in vigorous activity was low compared to the current study. Findings are comparable to the Chinese version MacFarlane et al. (2011) total PA ICC = 0.93 and moderate PA ICC 0.74, and Malaysian ICC total PA = 0.92 (Chu & Moy, 2015).

The CCC was high for total sitting time, while leisure time PA did not have the highest reliability unlike other studies. Studies claim that leisure PA is usually planned and is expected to have higher reliability when compared to less structured forms of PA (Sebastião et al., 2012). In this study a high level of CCC were obtained in housework/gardening. This might be explained by high levels of housework/gardening PA within the studied population compared to other studies 52.8 MET min per week (MacFarlane et al., 2011),834 MET min per week (Chu & Moy, 2015) and 597 MET min per week (Oyeyemi et al., 2014). The Cohen's Kappa obtained from this study was comparable to the Spanish IPAQ-long K = 0.61(Roman-Viñas et al., 2010).

The translation process of the MT-IPAQ-long was undertaken using four expert translators and cognitive interviews were carried out with different participants. The use of cognitive interviews was an added benefit of this translation. Since recall of PA is a complex cognitive process this ensured that the translation was accurate. Other translation studies limited the translation process to forward and back translation with expert or committee review (Kalvenas et al., 2016; Pedišić et al., 2011). The changes carried out due to the interview results improved the translation.

One of the limitations is that the MT-IPAQ-long is not yet validated. Appropriate reliability does not necessarily mean the tool is valid. A systematic review of the reliability and validity of different PA questionnaires found that most have similar concurrent validity when compared with an accelerometer (Helmerhorst et al., 2012). However, differences might arise as self-reported measures are based on perceived exertion, whilst accelerometers measure fixed values (Loney et al., 2011). Therefore a further study would be needed to establish the validity of the MT-IPAQ-long but we are confident in its validity due to extensive testing of the English and other versions.

A further limitation relates to the study sampling. The study population was recruited via convenience sampling which does not represent the Maltese population. As the questionnaire was self-administered it was expected that participants would have higher education levels. The sample included participants with different educational backgrounds and from different age groups. During cognitive interviewing a sampling bias towards participants with lower education was carried out ensuring that the questionnaire was understood by participants with lower education.

The results from this study highlights that the translated version of the MT-IPAQ-long is reliable. Device-based measures might be considered as being more accurate in the measurement of PA. Given the advances since their initial usage in the 1980s, device-based measures are becoming more affordable and are being used in large scale cohort studies, as well as, surveillance studies. However, as Troiano et al., (2014) highlighted, self-reported and device-based measure different aspects of PA. Combining self-reported and device-based measures for large scale studies is recommended for PA studies (Steene-johannessen et al., 2018). The use of MT-IPAQ long in conjunction with other device-base methods could be assessed in future studies to check for ability to measure changes in PA behaviour across different domains. Combining self-reported methods in combination with devices-based methods would result in the advocated paradigm shift (Troiano et al., 2012).

When assessing PA behaviours in Maltese speaking populations the translated version of MT-IPAQ-long will be an additional tool which PA researchers can use. The questionnaire will be promoted locally to public health specialists and researchers in the PA and health field. This final translated MT-IPAQ-long will be made available on the IPAQ website.

Conclusion

From this study it can be concluded that MT-IPAQ-long has reliability similar to that achieved in other IPAQ-long language translations and its reliability is comparable to that demonstrated in the original English language version. The results support the use of this tool for studies in healthy people among Maltese speaking populations. The translated Maltese version is a reliable tool and can be used with Maltese speaking individuals when measuring physical activity and sedentary behaviour.

Conflict of interest

The authors declare no potential conflict of interest.

Acknowledgements

We would like to thank the translators for their time, the experts for reviewing the translation, participants in the cognitive interviewing and those taking part in the questionnaire.

Funding

The research work disclosed in this publication is partly funded by the ENDEAVOUR scholarship scheme – Group B- National funds - Malta

References

Altman, D. G., & Bland, J. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *i*(fig 1), 307–310. Retrieved from http://www.sciencedirect.com/science/article/pii/S0140673686908378

Bauman, A., Ainsworth, B. E., Bull, F., Craig, C. L., Hagströmer, M., Sallis, J. F., … Sallis JF, Pratt M, S. M. (2009). Progress and pitfalls in the use of the international physical activity questionnaire (IPAQ) for adult physical activity surveillance. *Journal of Physical Activity and Health*, *6*(SUPPL. 1), 5–8. Retrieved from http://journals.humankinetics.com/AcuCustom/Sitename/Documents/DocumentItem/17424.pdf

Behling, O., & Law, K. S. (2011). *Translating Questionnaires and Other Research Instruments*. California: SAGE Publications, Inc.

Bergman, P. (2018). The number of repeated observations needed to estimate the habitual physical activity of an individual to a given level of precision. *Plos One*, *13*(2), e0192117–e0192117. https://doi.org/10.1371/journal.pone.0192117

Bergman, P., & Hagströmer, M. (2020). No one accelerometer-based physical activity data collection protocol can fit all research questions. *BMC Medical Research Methodology*, *20*(1), 1–8. https://doi.org/10.1186/s12874-020-01026-7

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy*, *86*(2), 94–99.

Chu, A. H. Y., & Moy, F. M. (2015). Reliability and validity of the malay international physical activity questionnaire (IPAQ-M)among a malay population in Malaysia. *Asia-Pacific Journal of Public Health*, *27*(2), NP2381–NP2389. https://doi.org/10.1177/1010539512444120

Craig, C. L., Marshall, A. L., Sjöström, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., … Oja, P. (2003). International physical activity questionnaire: 12-Country reliability and validity. *Medicine and Science in Sports and Exercise*, *35*(8), 1381–1395. https://doi.org/10.1249/01.MSS.0000078924.61453.FB

Doma, K., Speyer, R., Leicht, A. S., & Cordier, R. (2017). Comparison of psychometric properties between usual-week and past-week self-reported physical activity questionnaires: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *14*(1). https://doi.org/10.1186/s12966-017-0470-6

Dowd, K. P., Szeklicki, R., Minetto, M. A., Murphy, M. H., Polito, A., Ghigo, E., … Donnelly, A. E. (2018). *A systematic literature review of reviews on techniques for physical activity measurement in adults: A DEDIPAC study*. *International Journal of Behavioral Nutrition and Physical Activity* (Vol. 15). International Journal of Behavioral Nutrition and Physical Activity. https://doi.org/10.1186/s12966-017-0636-2

Grey, D. (2015). *Conducting Cognitive Interviews*. SAGE Publications Ltd. https://doi.org/https://dx.doi.org/10.4135/9781473910102.n6

Gropper, H., John, J. M., Sudeck, G., & Thiel, A. (2020). The impact of life events and transitions on physical activity: A scoping review. *PloS One*, *15*(6), e0234794. https://doi.org/10.1371/journal.pone.0234794

Hallal, P. C., Gomez, L. F., Parra, D. C., Lobelo, F., Mosquera, J., Florindo, A. a, … Sarmiento, O. L. (2010). Lessons learned after 10 years of IPAQ use in Brazil and Colombia. *Journal of Physical Activity & Health*, *7 Suppl 2*(Suppl 2), S259–S264.

Helmerhorst, H. J. F., Brage, S., Warren, J., Besson, H., & Ekelund, U. (2012). A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *The International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 103. https://doi.org/10.1186/1479-5868-9-103

Helou, K., El Helou, N., Mahfouz, M., Mahfouz, Y., Salameh, P., & Harmouche-Karaki, M. (2017). Validity and reliability of an adapted Arabic version of the long international physical activity questionnaire. *BMC Public Health*, *18*(1), 1–8. https://doi.org/10.1186/s12889-017-4599-7

Henrica, C. W. de. V. (2011). *Measurement in medicine a practical guide* (first). Cambridge: Cambridge University Press.

IPAQ group. (2019). IPAQ. Retrieved from https://sites.google.com/site/theipaq/questionnaire_links

Kalvenas, A., Burlacu, I., & Abu-Omar, K. (2016). Reliability and validity of the International Physical Activity Questionnaire in Lithuania. *Baltic Journal of Health and Physical Activity*, *8*(2), 29–41. https://doi.org/10.29359/bjhpa.08.2.03

Kelly, P., Fitzsimons, C., & Baker, G. (2016). Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered. *International Journal of Behavioral Nutrition and Physical Activity*, *13*(1), 1–10. https://doi.org/10.1186/s12966-016-0351-4

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661–671. https://doi.org/10.1016/j.ijnurstu.2011.01.016

Kwak, L., Hagströmer, M., & Sjostrom, M. (2012). Can the IPAQ-long be used to assess occupational physical activity? *Journal of Physical Activity and Health*, *9*(8), 1130–1137. https://doi.org/10.1123/jpah.9.8.1130

Lee, I. M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., Katzmarzyk, P. T., … Wells, J. C. (2012). Effect of physical inactivity on major non-communicable diseases worldwide:

An analysis of burden of disease and life expectancy. *The Lancet*, *380*(9838), 219–229. https://doi.org/10.1016/S0140-6736(12)61031-9

Liu, J., Tang, W., Chen, G., Lu, Y., Feng, C., & Tu, X. M. (2016). Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Archives of Psychiatry*, *28*(2), 115–120. https://doi.org/10.11919/j.issn.1002-0829.216045

Loney, T., Standage, M., Thompson, D., Sebire, S. J., & Cumming, S. (2011). Self-report vs. objectively assessed physical activity: Which is right for public health? *Journal of Physical Activity and Health*, *8*(1), 62–70. https://doi.org/10.1123/jpah.8.1.62

MacFarlane, D., Chan, A., & Cerin, E. (2011). Examining the validity and reliability of the Chinese version of the International Physical Activity Questionnaire, long form (IPAQ-LC). *Public Health Nutrition*, *14*(3), 443–450. https://doi.org/10.1017/S1368980010002806

Mannocci, A., Bontempi, C., Colamesta, V., Ferretti, F., Giraldi, G., Lombardi, A., … La Torre, G. (2014). Reliability of the telephone-administered international physical activity questionnaire in an Italian pilot sample. *Epidemiology Biostatistics and Public Health*, *11*(1), 1–9. https://doi.org/10.2427/8860

Milanovic, Z., PAntelic, S., Trajkovic, N., Jorgic, B., Sporis, G., & Bratic, M. (2014). Reliability of the serbian version of the International Physical Activity Questionnaire for older adults. *Clinical Intervention*, *4*(9), 580–587. https://doi.org/10.2147/CIA.S57379

Nordman, M., Matthiessen, J., Biltoft-Jensen, A., Ritz, C., & Hjorth, M. F. (2020). Weekly variation in diet and physical activity among 4-75-year-old Danes. *Public Health Nutrition*, *23*(8), 1350–1361. https://doi.org/10.1017/S1368980019003707

Oyeyemi, A. L., Bello, U. M., Philemon, S. T., Aliyu, H. N., Majidadi, R. W., & Oyeyemi, A. Y. (2014). Examining the reliability and validity of a modified version of the international physical activity questionnaire, long form (IPAQ-LF) in nigeria: A cross-sectional study. *BMJ Open*, *4*(12), 1–11. https://doi.org/10.1136/bmjopen-2014-005820

Pedišić, Ž., Jurakić, D., Rakovac, M., Hodak, D., & Dizdar, D. (2011). Reliability of the Croatian long version of the International Physical Activity Questionnaire. *Kinesiology*, *43*(2), 185–191.

Physical Activity Guidelines Advisory Committee. (2018). *Physical activity guidelines advisory committee scientific report*. Washington DC. https://doi.org/10.1111/j.1753-4887.2008.00136.x

Roman-Viñas, B., Serra-Majem, L., Hagströmer, M., Ribas-Barba, L., Sjöström, M., & Segura-Cardona, R. (2010). International Physical Activity Questionnaire: Reliability and validity in a Spanish population. *European Journal of Sport Science*, *10*(5), 297–304. https://doi.org/10.1080/17461390903426667

Rütten, A., Vuillemin, A., Ooijendijk, W., Schena, F., Sjöström, M., Stahl, T., … Ziemainz, H. (2003). Physical activity monitoring in Europe. The European Physical Activity Surveillance System (EUPASS) approach and indicator testing. *Public Health Nutrition*, *6*(4), 377–384.

Sebastião, E., Gobbi, S., Chodzko-Zajko, W., Schwingel, A., Papini, C. B., Nakamura, P. M., … Kokubun, E. (2012). The International Physical Activity Questionnaire-long form overestimates self-reported physical activity of Brazilian adults. *Public Health*, *126*(11), 967–975. https://doi.org/10.1016/j.puhe.2012.07.004

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), 1–10. https://doi.org/10.1136/bmjopen-2015-008430

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268.

https://doi.org/10.1093/ptj/85.3.257

Sjöström, M., Oja, P., Hagströmer, M., Smith, B. J., & Bauman, A. (2006). Health-enhancing physical activity across European Union countries: the Eurobarometer study. *Journal of Public Health*, *14*(5), 291–300.

Steene-johannessen, J., Anderssen, S. A., Ploeg, H. P. Van Der, Hendriksen, I. J. M., Donnelly, A. E., Brage, S., & Ekelund, U. (2018). Are Self-report Measures Able to Define Individuals as Physically Active or Inactive ? *Med Sci Sports Exerc.*, *48*(2), 235–244. https://doi.org/10.1249/MSS.0000000000000760.Are

Streiner, D., Norman, G. R., & Cairne, J. (2015). *Health measurment scales a practical guide to their development and use* (Fifth). Oxford: Oxford University Press.

Thompson, D. J. M., & Wesolowski, B. (2018). Standard Error of Measurement. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 1588–1590). Thousand Oaks: SAGE Publications, Inc. https://doi.org/10.1007/978-94-007-0753-5_2847

Troiano, R. P., McClain, J. J., Brychta, R. J., & Chen, K. (2014). Evolution of accelerometer methods for physical activity research. *Br J Sports Med*, *48*(13), 1019–1023. https://doi.org/10.1038/jid.2014.371

Troiano, R. P., Pettee Gabriel, K. K., Welk, G. J., Owen, N., & Sternfeld, B. (2012). Reported Physical Activity and Sedentary Behavior: Why Do You Ask? *Journal of Physical Activity and Health*, *9*(suppl 1), S68–S75. https://doi.org/https://doi.org/10.1123/jpah.9.s1.s68

UK Chief Medical Officers. (2019). *UK Chief Medical Officers ' Physical Activity Guidelines*. Retrieved from https://www.gov.uk/government/publications/physical-activity-guidelines-uk-chief-medical-officers-report

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, *17*(1), 101–110.

https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-

E