Factors Affecting Reliability of Grip Strength Measurements in Middle Aged and Older Adults [version 1; peer review: 2 approved with reservations]

Hugh Nolan¹, John D. O'Connor¹, Orna A. Donoghue¹, George M. Savva², Neil O'Leary³, Rose-Anne Kenny¹

¹The Irish Longitudinal Study on Ageing, University of Dublin, Trinity College, Dublin, Dublin, Ireland
²Quadram Institute Bioscience, Norwich, UK
³Health Research Board Clinical Research Facility, National University of Ireland Galway, Galway, Ireland

Abstract

Background: Grip strength is a well-established marker of frailty and a good predictor of mortality that has been measured in a diverse range of samples including many population studies. The reliability of grip strength measurement in longitudinal studies is not well understood.

Methods: Participants (n=130) completed a baseline and repeat health assessment in the Irish Longitudinal Study on Ageing. Grip strength was assessed using dominant and non-dominant hands (two trials on each). Repeat assessments were conducted 1-4 months later and participants were randomised into groups so that 50% changed time (morning or afternoon assessment) and 50% changed assessor between assessments. Intra-class correlation (ICC) and minimum detectable change (MDC₉⁵) were calculated and the effects of repeat assessment, time of day and assessor were determined.

Results: Aggregated measures had little variation by repeat assessment or time of day; however, there was a significant effect of assessor (up to 2 kg depending on the measure used). Reliability between assessments was good (ICC>0.9) while MDC₉⁵ ranged from 5.59–7.96 kg. Non-aggregated measures alone, taken on the non-dominant hand were susceptible to repeat assessment, time of day, assessor and repeated measures within-assessment effects whereas the dominant hand was only affected by assessor.

Conclusions: Mean and maximum grip strength had a higher ICC and lower MDC₉⁵ than measures on the dominant or non-dominant hands alone. The MDC₉⁵ is less than 8 kg regardless of the specific measure reported. However, changing assessor further increases variability, highlighting the need for comprehensive assessor training and avoiding changes within studies where possible.
Keywords
Grip strength reliability, Physical function, Longitudinal aging

This article is included in the TILDA gateway.

Corresponding author: John D. O'Connor (john.oconnor@tcd.ie)
Author roles: Nolan H: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation; O'Connor JD: Investigation, Methodology, Visualization, Writing – Review & Editing; Donoghue OA: Conceptualization, Investigation, Methodology, Project Administration, Writing – Review & Editing; Savva GM: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; O'Leary N: Data Curation, Formal Analysis, Investigation, Methodology, Writing – Review & Editing; Kenny RA: Conceptualization, Funding Acquisition, Investigation, Methodology, Resources, Supervision, Writing – Review & Editing
Competing interests: No competing interests were disclosed.
Grant information: Health Research Board Ireland [TILSA-2017-1]. TILDA is also funded by Atlantic Philanthropies, the Irish Government Department of Health and Children and Irish Life and Permanent plc. The SHARE-TILDA project was funded by the National Institute of Aging (Prime Award Number R21AG040387).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Copyright: © 2020 Nolan H et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
How to cite this article: Nolan H, O'Connor JD, Donoghue OA et al. Factors Affecting Reliability of Grip Strength Measurements in Middle Aged and Older Adults [version 1; peer review: 2 approved with reservations] HRB Open Research 2020, 3:32 https://doi.org/10.12688/hrbopenres.13064.1
First published: 03 Jun 2020, 3:32 https://doi.org/10.12688/hrbopenres.13064.1
Introduction
The use of maximum grip strength as a measure of muscle function and a proxy of overall body strength has become commonplace in clinical and epidemiological research. Grip strength deficits are used in models of frailty\(^1\,\^2\) and are predictive of fractures, disability, cognitive decline and mortality\(^3\,\)\(^-\)\(^5\). Grip strength is measured routinely in studies of aging\(^6\,\)\(^-\)\(^9\) where large numbers of participants necessitates quick, easily obtained and meaningful measurements. Clinically, grip strength is an important component in the assessment of sarcopenia recommended by the European working group on sarcopenia in older people\(^1\,\)\(^8\).

Grip strength is most often measured with a hand-held hydraulic dynamometer, which is inexpensive and portable. Several studies have examined grip strength test-retest reliability over time, positioning and assessor\(^1\,\)\(^9\)\(^-\)\(^14\). A recent review, revealed a lack of standardization in epidemiological studies which inhibits comparability of results\(^6\). Studies focusing on grip strength repeatability have stringent controls on assessment location and assessor types, and tend to have relatively small sample sizes. While these studies provide useful guidance, the real-world use of grip strength in epidemiological or clinical test batteries can be considerably different. In epidemiological studies, factors such as assessors, test environments and test order may not be strictly comparable to existing reliability studies and these factors may also change over time within the same study (e.g. see wave-to-wave changes in SHARE, ELSA, HRS).

In longitudinal studies, the ability to detect change over a long period of time is related to a measurement’s inherent variability, which is a function of the measurement tool, the testing paradigm, as well as the day-to-day human variability caused by transient factors that are generally not of interest to these studies. Characterising this variability in the short and medium term allows us to distinguish between a genuine change in performance and measurement error. It also allows us to identify improvements in protocols to ensure best results.

The aim of this study was to estimate the reliability of grip strength measures, and to test the effects of repeat assessments, assessor and time of day on grip strength reliability using a representative sample of Irish adults aged 50 years and older. We report the intra-class correlation (ICC), limits of agreement and the minimum detectable change (MDC), quantify variability both within-assessment and between-assessments, and investigate the changes associated with using the dominant or non-dominant hand.

Methods
Participants
Participants were recruited from the Survey of Health, Ageing and Retirement in Europe (SHARE), a cross-European longitudinal study constructed to enable multinational comparison on factors affecting ageing\(^10\,\)\(^-\)\(^17\). Within Ireland, SHARE recruited 1,119 community-dwelling adults aged 50 years and who participated in a baseline interview in 2006/2007. In 2010, the 827 remaining SHARE-Ireland participants were contacted and invited to participate in a detailed health assessment carried out in conjunction with The Irish Longitudinal Study on Ageing (TILDA), a large ongoing study epidemiologic study of health and ageing that is independent of SHARE\(^11\,\)\(^-\)\(^16\). Further information was provided to 377 participants, of whom 253 attended an initial health assessment. Participants reporting pain, recent surgery or injury were excluded. Ethical approval for this study was obtained from the Trinity College Dublin research ethics committee.

Health assessments
The health assessment followed similar protocols to those used in TILDA\(^11\) and took place within the dedicated TILDA health assessment centre at Trinity College Dublin. It consisted of a 3-hour battery of tests assessing anthropometric, cognitive, cardiovascular, gait, and visual function including a measure of hand-grip strength. All assessments were delivered in the same testing rooms using the same equipment by two trained research nurses, each with experience of delivering over 300 assessments in TILDA.

Grip strength measurement protocol
Grip strength was measured using a Baseline® hydraulic hand dynamometer. Two measurements were taken on both the dominant and non-dominant hands. Participants were asked to squeeze the dynamometer as hard as possible for a few seconds, while standing with their upper arm against their trunk and the elbow at 90 degrees. Participants who were unable to maintain this position could sit down or support the dynamometer with their free hand or a table. Mean and maximum grip strength was obtained across all four measurements and for dominant and non-dominant hand separately. To estimate within-participant variation, 130 participants attended a repeat health assessment in which the following factors were varied:

- Time between assessments varied from approximately 1–4 months (median 88 days; range 28–141 days; interquartile range 70–104 days).
- Time of day (morning/afternoon): Assessments were conducted in either the morning or afternoon, 50% of participants completed the repeat assessment at a different time of day to their initial assessment.
- Assessor: 50% of participants changed nurse at the repeat assessment.

Assessment lag, time of day and assessor was randomised using a minimisation routine designed to achieve balance between all combinations of these covariates, age group and sex of the participants.

Statistical analysis
The statistical analysis plan was based on previous works reported on reliability of cognitive and cardiovascular measures from the same sample\(^19\,\)\(^-\)\(^20\). The 95% limits of agreement were calculated for mean and maximum grip strength between the first and repeat assessment. The differences were visualised.
using Bland-Altman plots which graph the average of the two visits on the x-axis and the difference on the y-axis. In order to assess the factors affecting repeatability and calculate p-values, two mixed-effects multi-level models were fitted using Stata 15 (StataCorp LLC, TX, USA). Fixed effects were used to estimate factor treatments, and random effects to estimate variance contributions within assessments (the residual), between assessments for the same participant and between-participants. For the fixed effect reports in Table 1, models also included demographic factors (age, sex and height), to correct for potential imbalances in case distribution between assessors and between days. As this was only necessary for fixed effects, intercepts and random effects reported do not include these corrections. The ICC, which is an indicator of agreement within the random effect of the models, between repeat and baseline measurements was calculated for the groups in each model to indicate reliability across time.

\[
ICC = \frac{SD_{\text{within}}}{SD_{\text{between}}^2 + SD_{\text{within}}^2}
\]

Where SD\textsubscript{Between} is the between-individual standard deviation and SD\textsubscript{within} is within-individual standard deviation. The MDC\textsubscript{95}, which indicates the value below which 95% of change scores are likely to lie if measurement error alone accounted for them, was calculated for each measure by:

\[
MDC_{95} = \sqrt{2} \times 1.96 \times SD_{\text{within}}
\]

For all models, confidence intervals for the ICC and MDC\textsubscript{95} values were calculated using bootstrapping with 1000 repetitions. Bootstrapping was clustered by participant and, in order to mimic sampling processes, was stratified by 1) lag between assessments (less than/greater than median), 2) assessor (change/no change at repeat assessment), and 3) time of day (change/no change at repeat assessment).

Model 1: Estimating the reliability of aggregated measures between assessments
Model 1 was used to model the effects of participant and setting factors on mean and maximum grip strength measures. Mean and maximum were calculated across all four measurements (first and second measurements for dominant and non-dominant hands) for each assessment. Fixed effects included assessment (repeat/baseline), lag between assessments (in days), time of day (morning/afternoon) and assessor.

### Table 1. Results from mixed effects models to assess reliability of aggregated measures of grip strength (model 1; mean grip and max grip columns) and individual measures of grip strength (model 2; dominant hand, non-aggregated grip and non-dominant hand, non-aggregated grip). Coefficients for the fixed part of the model represent the influence of each condition, e.g. repeat measurement for mean grip was 0.47 kg (-0.07–1.01) higher than initial. The coefficient for lag between assessments is per day. The random effects represent the variability between participants, assessments and within assessment. Intercept values are the mean when fixed effects are zero.

<table>
<thead>
<tr>
<th></th>
<th>Mean grip</th>
<th>Max grip</th>
<th>Dominant hand non-aggregated grip</th>
<th>Non-dominant non-aggregated grip</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>30.70 (24.18 – 37.21)</td>
<td>32.24 (25.46 – 39.02)</td>
<td>31.89 (23.82 – 38.95)</td>
<td>31.15 (24.61 – 37.70)</td>
</tr>
<tr>
<td><strong>Fixed Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between assessments, repeat vs initial (kg)</td>
<td>0.47 (-0.07 – 1.01)</td>
<td>0.32 (-0.21 – 0.85)</td>
<td>0.25 (-0.45 – 0.95)</td>
<td>0.71* (0.14 – 1.27)</td>
</tr>
<tr>
<td>Time of day, a.m. vs p.m. (kg)</td>
<td>-0.72 (-1.50 – 0.05)</td>
<td>-0.72 (-1.49 – 0.04)</td>
<td>-0.85 (-1.81 – 0.11)</td>
<td>-0.83* (-1.64 – -0.02)</td>
</tr>
<tr>
<td>Assessor (kg)</td>
<td>1.72*** (0.96 – 2.49)</td>
<td>1.10** (0.34 – 1.86)</td>
<td>1.95** (0.99 – 2.92)</td>
<td>1.16** (0.38 – 1.94)</td>
</tr>
<tr>
<td>Lag between assessments (days)</td>
<td>0.02 (-1.05 – 1.10)</td>
<td>0.02 (-1.03 – 1.10)</td>
<td>-0.13 (-1.39 – 1.12)</td>
<td>-0.13 (-1.17 – 0.92)</td>
</tr>
<tr>
<td>Within-assessment, repeat vs initial (kg)</td>
<td>NA</td>
<td>NA</td>
<td>-0.17 (-0.53 – 0.20)</td>
<td>-0.87*** (-1.21 – -0.53)</td>
</tr>
<tr>
<td><strong>Random Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.D. Between participants (kg)</td>
<td>9.57 (8.39 – 10.9)</td>
<td>9.94 (8.73 – 11.3)</td>
<td>10.1 (8.81 – 11.5)</td>
<td>9.35 (8.2 – 10.7)</td>
</tr>
<tr>
<td>S.D. Between assessments (kg)</td>
<td>2.04 (1.79 – 2.33)</td>
<td>2.02 (1.76 – 2.3)</td>
<td>2.11 (1.74 – 2.58)</td>
<td>1.65 (1.32 – 2.06)</td>
</tr>
<tr>
<td>S.D. Within assessment (kg)</td>
<td>NA</td>
<td>NA</td>
<td>1.94 (1.77 – 2.13)</td>
<td>1.81 (1.65 – 1.99)</td>
</tr>
<tr>
<td>ICC within-participant, within-assessment</td>
<td>NA</td>
<td>NA</td>
<td>0.966 (0.957 – 0.974)</td>
<td>0.967 (0.961 – 0.974)</td>
</tr>
<tr>
<td>ICC – between assessment</td>
<td>0.956 (0.946 – 0.967)</td>
<td>0.961 (0.95 – 0.97)</td>
<td>0.925 (0.902 – 0.947)</td>
<td>0.933 (0.917 – 0.948)</td>
</tr>
<tr>
<td>MDC\textsubscript{95} – between assessment (kg)</td>
<td>5.66 (4.99 – 6.33)</td>
<td>5.59 (4.97 – 6.20)</td>
<td>7.96 (7.05 – 8.86)</td>
<td>6.80 (6.25 – 7.34)</td>
</tr>
</tbody>
</table>

*p<0.05 **p<0.01 ***p<0.001.
Corrections for fixed effects (age, sex and height) and random effects (participant) were also included. For this model, a single ICC and $\text{MDC}_{95}$ are calculated for the participant group (within-participant, between-assessment).

Model 2: Estimating the reliability of individual measures within assessments

While Model 1 uses aggregated grip strength measures (mean and maximum), Model 2 considers the effects of repeated measures within assessment and differences between dominant and non-dominant hands. Fixed effects included assessment (repeat/baseline), lag between assessments (in days), time of day (morning/afternoon), assessor, repeat within-assessment (first vs second) and hand (dominant/non-dominant). The interactions of hand with each of the other fixed effects were also included. Participant specific fixed effects (age, sex and height) and random effects (participant and assessment which was nested within participant) were also included. Random effects were estimated separately for dominant and non-dominant hands (nested random coefficient models). For this model, two ICCs were estimated: one for the reliability within-assessments, another between-assessments. $\text{MDC}_{95}$ is estimated for individual grip strength measures for both dominant and non-dominant hands.

Results

This analysis is based on 130 participants (median age 66 years, range 50–89 years; 55% female). Grip strength data was available at baseline and repeat assessments for 123 participants, with 21 of these having incomplete data at one or both the assessments due to injury (Figure 1). The 95% limits of agreement between baseline and repeat assessments were -6.2–7.0 kg for mean grip strength and -5.9–6.5 kg for maximum grip strength (Figure 2).

Mixed effects models

The results of the mixed effects models in Table 1 are separated into components associated with the fixed effects and random effects. For the fixed effects, a positive result indicates an increase in grip strength at the repeat assessment, in the morning, when the assessment was delivered by assessor 2, with every 1 month increase in the lag between assessments and in the second measure taken within the same assessment.

The $\text{MDC}_{95}$, ICC, intra-group, and inter-group standard deviations are listed for each model. All ICC point estimates were $>0.9$. The between-assessment ICC for dominant and non-dominant hands is typically between 0.92 and 0.93, while aggregate measure ICCs are about 0.96. $\text{MDC}_{95}$ for mean grip strength...
and maximum grip strength (aggregated from 2 measures of both dominant and non-dominant hands) is about 5.6 kg. For non-aggregated measures, MDC95 was 6.8 kg for the non-dominant hand and 7.96 kg for the dominant hand. There was little evidence of a difference in mean, maximum or dominant-hand grip strength between baseline and repeat assessments; however, there was some evidence of a difference between assessments for the non-dominant hand (difference=0.71 kg, p=0.014). Similarly, mean, maximum and dominant hand grip strength did not vary significantly with time of day but grip strength measured in the non-dominant hand was lower when assessments took place in the morning (difference=0.83 kg, p=0.043). Changing assessor affected overall mean and maximum grip strength and also individual measurements taken using the dominant and non-dominant hands (differences range from 1.16-1.95 kg). When two grip strength measures were taken using the non-dominant hand within the same assessment, it was higher in the first measure compared to the second measure (difference=0.87 kg, p<0.001).

**Discussion**

This study presents reliability estimates for mean, maximum, and individual grip strength measures in community-dwelling adults aged 50 years and over, along with breakdowns of the variance structure of the measurements and contributing factors to grip strength changes. Mean and maximum grip strength have good test-retest reliability with an ICC of 0.95 or greater between assessments up to 4 months apart. The minimum detectable change for an individual is approximately 5.6 kg, similar to MDC95 values of 4.7–6.2 kg reported previously. This represents a 17–18% change from the mean value for the two measures of grip strength. The overall 95% limits of agreement, which incorporates the MDC and fixed effects, are slightly narrower but similar to values in previous studies i.e. -8.3–7.2 kg and -7.19–8.75 kg. Studies assessing reliability differences between aggregated and non-aggregated grip measures have shown contrasting results with some suggesting a single measure has similar test-retest reliability to mean or maximum of several trials, while others recommend aggregation.
We found that using the mean or maximum grip strength gives a higher ICC and lower MDC \(_{95}\) compared to individual measurements from either the dominant or the non-dominant hand. This indicates that while reliability is high using both approaches, it will be more difficult to identify a genuine change in performance when using a non-aggregated measurement from one hand only. Although a range of studies and reviews have reported test-retest reliability \(^{9,10,13}\) and normative values for grip strength \(^{3,17}\), the present work focuses on its application in longitudinal studies of aging and provides minimum detectable change values to guide interpretation of longitudinal changes. As expected, results suggest aggregated measures should be used where possible with the benefit of decreasing the MDC \(_{95}\) from around 7-8 kg (non-aggregated measures) to 5.6 kg.

Differences in mean and maximum grip strength were small across repeat assessments and with varying lag times between assessments. There were differences between morning and afternoon measurements of up to 0.85 kg; however, only the non-dominant, non-aggregated measure was significant at an alpha level of 0.05. When examining non-aggregated grip strength measurements on the non-dominant hands only, performance was higher when it was obtained in the repeat assessment, in the morning, and in the first measure within an assessment. The magnitude of these changes is small, and there is little evidence to suggest either dominant or non-dominant hand is more affected by time of day. The interaction between hand used and repeated measures within-assessment suggests that the non-dominant hand may be more susceptible to fatigue when multiple measurements are taken on the same occasion.

The most striking result is the significant assessor effect of 1-2 kg observed across all grip strength measures (mean, maximum, dominant, non-dominant). Equipment and environmental factors confounding this effect were ruled out by careful experimental design, and as there is little overall difference between repeat assessments or time of day, we conclude that this effect is likely related to the nurse administering the test. The intensity of instruction, even using the same wording, can affect grip strength measurements \(^{9,10,13}\) and the positive consequences of effective encouragement are noted in other studies of exercise and physical activity \(^{9,11}\). Clinical studies have also found that the patient-clinician relationship affects healthcare outcomes \(^{97}\). Throughout the health assessment in this study (minimum 3 hours), it is likely that the nurse built a relationship with the participant. Regardless of the potential explanations, variation due to the assessor is a fixed effect which is additive to the reported MDC \(_{95}\) and would increase variance and the bounds of detectable change in cases where the assessor is not held constant. Additional training of assessors to raise awareness of the effects of wording and encouragement may improve repeatability of grip strength tests.

The differences observed here are calculated between only two nurses, each with considerable experience administering health assessments, and so it is likely that even larger variation would be observed among a larger group of assessors. This highlights the importance of comprehensive assessor training especially in large studies which require multiple assessors and where repeated measurements are common. It is often difficult to retain testers in long-term longitudinal studies; this does introduce a potential biasing factor, which should at least be recorded so that potential rater effects can be accounted for in analysis.

The strengths of this study include the strict study design which allows separation and quantification of the different sources of variation, and the relatively large number of participants. A limitation is that the health assessments were carried out by only two nurses. Most epidemiological studies would use a higher number of assessors: having more nurses would have allowed us to estimate a distribution of assessor offsets and to fully understand the impact of using multiple assessors.

**Conclusions**

Here we report the effects of time of day, assessor, and practice on measures of grip strength, along with their standard deviations within assessment, between assessments and between participants. We demonstrated that grip strength measurements are reliable when obtained at repeat assessments over 1-4 months and at different times of day but that they are potentially affected by different assessors. Grip strength measurements were more susceptible to repeat assessment, time of day and repeated measures effects when they were obtained using the non-dominant hand. We also derive estimates for MDC \(_{95}\) which can be used to assess changes in grip strength performance in individuals drawn from comparable populations. These results suggest that longitudinal studies should use aggregated grip strength measures as well as minimising the number of assessors over the course of the study.

**Data availability**

**Underlying data**

The data for this study is linked to the wider TILDA dataset containing sensitive, personal information and access is therefore granted on a case-by-case basis following application (including assurances that the data anonymity will be protected and a description of use) to The Irish Longitudinal Study on Ageing (email: tilda@tcd.ie).

**Acknowledgements**

The authors would like to thank the TILDA participants and staff for their ongoing commitment to the study. Author contributions: 1) Study concept and design - HN, OD, GS, NOL, RAK 2) Acquisition of data - HN, OD, RAK 3) Analysis and interpretation of data - HN, JOC, OD, GS, NOL, RAK 4) Drafting of the manuscript - HN, JOC, OD, GS, NOL, RAK 5) Critical revision of the manuscript for important intellectual content - HN, JOC, OD, GS, NOL, RAK.
This article on the reliability of grip strength measurements in longitudinal studies is of interest. There are numerous similar studies in the field of rehabilitation with generally well accepted standardization of measurement positions and instructions. This study adds to the body of knowledge in looking at factors affecting the measurement of grip strength in longitudinal studies led by clinicians who may not have the same knowledge and experience of usual grip strength measurement protocols. An earlier publication has identified that there was a lack of standardized procedures for grip strength measurement in epidemiological studies, but no one has actually assessed if this was affecting the reliability and validity of these measurements in this context. This is important considering the growing number of publications reporting on grip strength in various fields.

We think it is interesting to know that reliability of grip strength measurements is high in this study. The authors have highlighted that the most important source of error was related to differences in assessors and this stresses the importance of standardized procedures and training. This will hopefully help future studies in elaborating a better measurement protocol.

Our strongest criticism is related to the absence of description of the grip strength protocol used. This limits the generalizability of the data on measurement errors.

We understand that this is part of a set of studies which have looked at the error of measurements of other tests and this should be stated in the introduction to help the reader to understand the context of this study.

Below is a list of issues that would benefit from a revision to improve the quality of the manuscript.
**Introduction**
Second paragraph, second sentence: “A recent review revealed a lack of standardization in epidemiological studies which inhibits comparability of results”. Since this review was published in 2011, it would be appropriate to remove the term "recent".

The rationale of the study is not clear from the two following sentences:

“Studies focusing on grip strength repeatability have stringent controls on assessment location and assessor types, and tend to have relatively small sample sizes. While these studies provide useful guidance, the real-world use of grip strength in epidemiological or clinical test batteries can be considerably different”.

It is not clear why it can be considerably different in epidemiological studies and why stringent protocols should not be used in this type of study to ensure the validity of the measurements.

**Method**
The description of the test protocol is incomplete. If possible, the following should be described:

1. The number of dynamometers used. Did each evaluator use a different dynamometer?

2. Calibration verification of the dynamometer(s).

3. The handle and the arm positions.

4. The sequence of the measurements between the dominant and the non-dominant sides (Were the measurements taken consecutively on the same side?).

5. The sequence of the overall evaluation (Were the tests administered in the same order? If yes, were the grip strength measurements done at the beginning or the end?).

6. The training of the assessors.

7. The blinding of the assessors. As the authors wanted to assess the “real world” error of measurement, this would be an important factor to report.

8. The number of people who could not maintain the standing position or had support to hold the dynamometer? Why was it not chosen to perform the test sitting down and to support the dynamometer in the first place to ensure that everyone used the same position?

The values used in the analyses must be specified. The following sentence should be reworded to help clarity. “Mean and maximum grip strength was obtained across all four measurements and for dominant and non-dominant hand separately.”

It is quite unusual to use the mean of both left and right hands in grip strength studies. Whilst we understand why this was chosen in this type of study, it wasn't very clear that this was what happened.

It is not clear from the introduction or the method section why the authors chose the factors under consideration (time, assessors, etc.). Could it be justified why these factors were chosen in
relation to the literature on grip strength measurements?

Calculation of the minimum detectable change is usually made from the SEM. To make it easier for less experienced readers to understand, could you explain why you chose the SDwithin.

**Results**
The distribution of the participants age could be better estimated with the mean, standard deviation and range, or at least with the IQR.

Are there any health problems that might affect strength that would be of interest to report in the sample description? This would help to establish the generalizability of their results.

**Discussion**
In light of their results and the need for standardized grip strength measurements when using more than one evaluator, the authors could add a reference to published standardized protocols.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** AMB: Methodology, Assessment, Psychometrics; KF: Hand Therapy, Assessment, Psychometrics.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.
Annie Robitaille
Faculty of Health Sciences, University of Ottawa, Ottawa, Canada

The study entitled “Factors Affecting Reliability of Grip Strength Measurements in Middle Aged and Older Adults” examines the effect of repeat assessments, change in assessor, and time of day on grip strength reliability. Grip strength is frequently used in research studies and in the clinical setting and is often used as a proxy for frailty which makes the topic of the paper interesting and important for readers. However, there are some concerns that need to be addressed and there is already a good deal of research looking at the reliability of grip strength tests and so I was hoping the authors could further elaborate on the implications and potential impact of their findings.

Participants
- The participant section would need some clarifications. What does this sentence mean “Further information was provided to 377 participants, of whom 253 attended an initial health assessment.”? Do you not mean that further information was collected from 377 participants or do you mean that 377 were sent information asking them whether they would participate but only 253 completed the baseline assessment? Also, how many were excluded because of pain, recent surgery or injury? How do you then get to 130? Might also be good to refer to Figure 1 here.
- Remove the first-word study in “a large ongoing study epidemiologic study”
- Should “community-dwelling adults aged 50 year” be community-dwelling adults aged 50 years and older?
- I think it would also be important to elaborate on the training level of the 2 nurses who did the assessments since the authors mention how it is important to train nurses on grip strength assessment and how this can have an impact on scores.

Grip strength measurement protocol
- Please be more specific about the time between assessments. In my opinion, 1-4 months and the mean is not enough information. How many had an assessment within less than a month, how many had an assessment after a month but less than 2 months... Why was 1-4 months even chosen?
- Please be more specific about time of day of testing. I ask because 11:30am and 12:30pm is not all that different. This might explain the lack of significant results. Until more information is provided, it is hard to understand the actual importance of the findings. Also, these are 3-hour long testing sessions. When is the grip strength administered within these 3 hours? Are all tests always administered in the same order?
- What was the actual protocol used (e.g., instructions) to the participants? Please provide more information.

Results
○ What about age and gender? There is no mention of these results. Why wasn't occupation also included?

Discussion
○ Please elaborate on the statement “Equipment and environmental factors confounding this effect were ruled out by careful experimental design”. I would actually do so in the methods section.

○ I would also ask that the authors further comment on the potential importance of their findings. I think this is missing. For example, when the authors write “We demonstrated that grip strength measurements are reliable when obtained at repeat assessments over 1-4”, please elaborate on the implications. What are the implications for studies such as SHARE, ELSA, HRS... that collect data every 1-3 years? In my opinion, this is the biggest limitation of this study. Please make sure to highlight what is novel about your study compared to the previous literature and its impact on research knowledge and clinical practice.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Older adults, frailty, dementia, longitudinal data analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.