University of Groningen

Reliability of an instrument for screening hand profiles

Woldendorp, Kees H.; de Schipper, Antoine W.; Boonstra, Anne M.; van der Sluis, Corry K.; Arendzen, J. Hans; Reneman, Michiel F.

Published in:
JOURNAL OF HAND THERAPY

DOI:
[10.1016/j.jht.2018.05.002](https://doi.org/10.1016/j.jht.2018.05.002)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Final author's version (accepted by publisher, after peer review)

Publication date:
2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Woldendorp, K. H., de Schipper, A. W., Boonstra, A. M., van der Sluis, C. K., Arendzen, J. H., & Reneman, M. F. (2018). Reliability of an instrument for screening hand profiles: The Practical Hand Evaluation. *JOURNAL OF HAND THERAPY*, 31(4), 544-553. <https://doi.org/10.1016/j.jht.2018.05.002>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Contents lists available at ScienceDirect

Journal of Hand Therapy

journal homepage: www.jhandtherapy.org

Scientific/Clinical Article

Reliability of an instrument for screening hand profiles: The Practical Hand Evaluation

Kees H. Woldendorp MD^{a,*}, Antoine W. de Schipper BSc^b, Anne M. Boonstra PhD, MD^a,
Corry K. van der Sluis PhD, MD^c, J. Hans Arendzen PhD, MD^d, Michiel F. Reneman PhD^c

^a "Revalidatie Friesland" Center for Rehabilitation, Beetsterzwaag, The Netherlands

^b Hogeschool van Amsterdam, University of Applied Sciences, Amsterdam, The Netherlands

^c Department of Rehabilitation, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

^d Department of Rehabilitation Medicine, Leiden University Medical Center, Leiden, The Netherlands

ARTICLE INFO

Article history:

Received 14 January 2018

Received in revised form

20 April 2018

Accepted 9 May 2018

Available online xxx

JV

Keywords:

Screening instrument

Assessment

Upper extremity

Musculoskeletal complaints

Hand problems

ABSTRACT

Study Design: Psychometric study with 2-week interval.

Introduction: Musculoskeletal hand complaints are common among manual workers. Mismatch between anthropometric hand features and tasks can affect the ability to perform hand activities, with an increased risk of complaints. Although screening of these features may improve diagnosis and treatment, no validated screening tool is available. The Practical Hand Evaluation (PHE) screening tool might fill this gap, but its psychometric properties are unknown.

Purpose of the Study: To test the reliability of the PHE and to explore the feasibility of item reduction of the PHE.

Methods: Right-hand profiles of 117 healthy volunteers (66 women, 51 men; mean age, 22.8 years) were independently assessed 4 times by 6 couples of researchers using the PHE, twice on day 1 and twice 2–3 weeks later. Intrarater and inter-rater reliability (intraclass correlations), standard error of measurement (SEM), potential confounding factors (gender, joint hyperlaxity, and measurement order) affecting the instrument's reliability (limits of agreement), and collinearity between the PHE items were determined (variation inflation factor analysis and hierarchical clustering of correlation coefficients).

Results: The intrarater and inter-rater reliabilities of the PHE were good for 12 of 14 items (86%; $r = 0.67-0.90$). Absolute SEM varied between 2.01 and 9.23 mm. The percentage of shifts of at least 2 classes in a repeated measurement was <15%. Cluster analysis identified 6 clusters of hand items.

Discussion: The reliability for nearly all PHE items is good. Measurement errors were substantial relative to variances in the reference population, but not to gender, joint laxity and order of administration. Clustering into 6 separated clusters of items was possible.

Conclusions: The PHE fulfills many of the criteria for screening of anthropometrics of the hand. Its reliability is high. The SEM might be improved with future adaptations toward a digital photographic PHE. Reduction to 6 items seems also possible.

© 2018 Hanley & Belfus, an imprint of Elsevier Inc. All rights reserved.

Introduction

The lifetime prevalence of musculoskeletal problems of the upper extremities in professions involving intensive or repetitive manual work, including desktop workers, dental hygienists,^{1–4} professional sportsmen,^{5–7} and musicians, is high (up to 55%).^{8,9}

Conflict of interest: All named authors hereby declare that they have no conflicts of interest to disclose.

* Corresponding author. "Revalidatie Friesland" Center for Rehabilitation, P.O. Box 2, 9244 ZN Beetsterzwaag, The Netherlands. Tel.: +31 512 389295; fax: +31 512 389244.

E-mail address: k.h.woldendorp@revalidatie-friesland.nl (K.H. Woldendorp).

These problems frequently interfere seriously with work performance and even with careers.¹⁰ Much research has been performed among professional musicians because of the extreme physical demands musicianship may make on upper extremities, particularly on the hands.^{8,9,11,12} For example, studies have reported frequencies of bimanual finger movements up to 1800 per minute among pianists or extreme finger positions among violinists during many hours of exercise.¹³ If musicians have unfavorable anthropometric hand features, like hand and finger sizes,^{14–16} these high upper extremity work demands may amplify biomechanical stresses of the hand and fingers, and many hours of playing increase exposure time to this stress.¹⁴ Relationships have been reported between anthropometric features of the hand and

playing-related pain.^{17–22} However, 2 systematic reviews among musicians^{12,23} found no consistent results regarding these features, which might be due to the diversity of anthropometric aspects they included.

Screening of the anthropometric features of the manual workers' hands would seem to be a relevant step in the diagnostic process for hand problems in various populations of manual workers. This screening method should be valid, reliable, unambiguous, and not time consuming. One such clinical assessment tool for the screening of anthropometric features of musicians' hands has been described by Wagner.¹⁴ This tool, the Practical Hand Evaluation (PHE), consists of 14 items regarding the size of the hand and the distances between the fingertips, which can be measured with ruler, pencil, and paper within 10 minutes. The PHE has been used to calculate percentiles for the individual anthropometric hand characteristics from a large sample of professional German musicians. This yielded associations between musculoskeletal hand complaints (MSHCs) and certain hand measurements, such as a small span between thumb and little finger in pianists.¹⁵

However, none of the psychometric values of the PHE have been established, for example, the intrarater and inter-rater reliability and standard error of measurement (SEM), as these are priorities when establishing the clinical utility of any instrument. The goal of this study was therefore to determine if the PHE is a reliable screening instrument for the anthropometric features of the hand. Moreover, we assumed that assessing all 14 items of the PHE might yield redundant information about the anthropometric values of the hand, and reducing the number of items might enhance its clinical usability as a screening instrument. We therefore examined whether reduction of items was statistically feasible. We also studied the influence of potential confounding factors on the reliability of the PHE.

Subsidiary goals were as follows:

1. Testing the intrarater and inter-rater reliability for the PHE items.
2. Testing the measurement error and the potential impact of this measurement error on the classification of PHE items.
3. Testing the influence of gender, joint hyperlaxity, and measurement order on the reliability.
4. Exploring the feasibility of reducing the number of items in the PHE.

Method

Study design and participants

We conducted a psychometric study. The study used 2 types of participants; rater participants and subject participants (from whom the hand features were measured). Subject participants included third-year physical therapy students and staff of the Hanze University of Applied Sciences in Groningen (the

Netherlands), recruited in 2016 and 2017. Information was provided via an explanatory letter, with additional oral information as needed. Subject participants were eligible if they were 17–50 years; had no self-reported pathology, disorders, or complaints of the right hand; and were able to fill out the questionnaire (English or Dutch version). All subjects gave informed consent. The medical ethics committee (METc 2016/220) decided that no formal approval was required. The guidelines for reporting reliability and agreement studies were used for conducting and reporting of the study.²⁴

Procedure

The test conditions were kept as constant as possible, including room temperature and acclimatization of the participants for at least half an hour before test administration to minimize the influence of unwarranted sources of bias. The raters received training from one of the authors (KHW, with appr.20 years of experience as a hand rehabilitation physician), with approximately 20 years of experience as a hand rehabilitation physician, to optimize the consistency of the test procedure. Training consisted of two 2-hour sessions for practical instruction on the assessment procedure, exercising, and receiving feedback until the raters had mastered the procedure. Tests were independently administered by 6 fixed couples of raters. The right hand of each of the subject participants was assessed in 4 measurements on 2 days by a rater couple (referred later as rater A and rater B). On the first day, the subject participant was measured by each of the 2 raters, with an interval of approximately half an hour, to calculate inter-rater reliability at short intervals (rater A – 30 minutes – rater B) (Fig. 1). Two to three weeks later, subjects were measured again by the same rater couple to establish intrarater and inter-rater reliability at long intervals (Fig. 1). The raters were blinded for their own and other raters' results.

On day 1, the procedure started by having each subject participant complete a short questionnaire to check the eligibility criteria for the study and to collect information on language, age, gender, hand dominance, and body height. Joint hyperlaxity of each participant was assessed to determine its possible influence on the reliability of the PHE (see later). This was done on day 1 by the first rater. Subsequently, the hand characteristics were assessed: participants were sitting in a stable but relaxed position behind a table in a quiet room, with rater A or B sitting across the same table. Rater A or B marked the center of the participants' fingertips with removable ink to enable optimal measurement. Raters A and B simultaneously measured 2 subject participants in different rooms (round 1). Half an hour after the assessment, the subject participants were retested in the other room by the other researcher (round 2). All marking points on the fingertips were removed before the subject participants entered the next room. Subject participants were instructed not to correct the other rater if they noticed differences in test performance. After 2–3 weeks, the procedure was repeated with the same rater couple for each subject

Fig. 1. (A, B) Schedule of measurements by the 6 rater couples.

participant. Before the second assessment, subject participants completed a short questionnaire to assess if their situation had been stable since the previous measurements, and if there were reasons for exclusion from the retest due to hand problems. The orders of measurement for 1 rater couple on the 2 days were as follows: ABAB, BABA, ABBA, or BAAB. A percentage of the subject participants were tested using either the same order of items (for the particular rater couple) or in reversed order to determine the influence on the measurement error.

The 12 raters involved measured 2 groups, 1 in 2016 and 1 in 2017. This was done because there were actually 2 consecutive studies with partially different study goals. Because the 6 raters of the second study tested the PHE items against another (photographic) assessment procedure, they assessed only 4 of the 14 hand-related items of the PHE (span 1-5, span 2-5, difference 1-3, and hand length). We assumed that these items were representative of the clusters of PHE items that might be most clinically relevant in screening hand anthropometrics. The other 10 PHE items were not included in the second study, resulting in a smaller number of participants for testing the reliability values of these 10 items (Table 1).

Measures

Practical Hand Evaluation

Wagner¹⁴ developed a standardized assessment procedure for hand features by measuring 3 different biomechanical components: shape (5 items), passive range of joint motion (15 items), and muscle strength (15 items). For the purpose of clinical screening, Wagner reduced the number of items from 35 to 14 to create the PHE (Appendix A; Fig. 2).

Joint laxity

Joint laxity was assessed twice on day 1 by both raters independently by applying the 5 criteria developed by Beighton.²⁵ The Beighton score is a simple 9-point system to quantify joint laxity (score, 0-9); higher scores mean more laxity, and a score of 4 or higher is interpreted as hyperlaxity.²⁵ One point is scored if the subject can place their palms on the ground (with legs straight) while standing bent forward, 1 point is given for each elbow and each knee that bends backward, 1 point for each thumb that touches the forearm when bent backward, and 1 point for each little finger that bends backward for more than 90°.

Data analysis

Demographic characteristics are presented as mean and standard deviation (SD; for interval/ratio data) or percentages (for dichotomized data). The percentage of missing data among the PHE items was calculated over the total of measurements per item: percentages $\leq 3\%$ were judged as acceptable, between 3%

and 15% as doubtful, and $>15\%$ as not acceptable.²⁶ Missing data were handled via pairwise deletion; a particular case was used if it had a missing value when analyzing other variables with non-missing values.

Reliability

Reliability was assessed using 1-way random intraclass correlation coefficients (ICCs) because of the multiple raters, with 95% confidence intervals. Intrarater and inter-rater reliabilities were analyzed using ICC for continuous data if data were normally distributed.²⁷ Normality of data distribution was checked visually via histograms for each of the 14 different PHE items and also for these items when calculated for each of the 3 factors separately. Spearman rho and ICC were both determined if data were not normally distributed. More specifically, we used the ICC 1-way random-effects absolute agreement (single measurement) for the inter-rater reliability because the different rater couples rated different participants (Table 1). For intrarater reliability, we opted for the ICC 2-way mixed-effects absolute agreement (single measurement).²⁸ Interpretation of the ICC for reliability measures was as follows: <0.50 = poor, $0.50 \leq x < 0.75$ = moderate, $0.75 \leq x < 0.90$ = good, and ≥ 0.90 = excellent.²⁹ Spearman's rho for the items with non-normal distribution differed very little (<0.02) from the ICCs. For reasons of clarity, only ICCs are presented here. Inter-rater reliability_{short} (within day 1; assessment moment 1-2) and inter-rater reliability_{long} (for more than 2 days; assessment moment 1-3 or 1-4) were both determined.

Measurement error

The measurement error between 2 raters on 2 different days was calculated in 2 ways: the absolute error of measurement (SEM) and the shift percentage of the measurement error (percentage of shifts). The latter represents the chance that a particular PHE item of a subject participant will end up in a different category when rated by the second rater. SEM and coefficient of variation were assessed using $SD \cdot \sqrt{1 - \text{inter-rater reliability}_{\text{short}}}$ ³⁰ and SD/mean , respectively. Percentage of shifts was estimated in 3 steps. First, the percentile data of the sample used by Wagner¹⁴ were divided into 7 categories: extremely small if $x \leq P5$; very small if $P5 < x \leq P10$; small if $P10 < x \leq P30$; average if $P30 < x \leq P70$; large if $P70 < x \leq P90$; very large if $P90 < x < P95$; and extremely large if $x \geq P95$. Second, test scores were converted into these categories based on the reference tables provided by Wagner.¹⁴ Third, we calculated by how many classes each subject had shifted. Finally, we determined the percentage with a shift of ≥ 2 categories. We assumed that a shift of ≥ 2 of 7 percentile categories was a relevant shift in percentile categories erroring clinical practice. We defined a shift percentage of $<5\%$ as acceptable, one of $5\% \leq x < 10\%$ as substantial and one of $\geq 10\%$ as large.

Table 1
Schedule of rater couples and the characteristics of the ratings

Rater couple	Period	Subjects	N	Item order		Order raters		Items
				N down	N up	N ABAB/BABA	N ABBA/BAAB	
I	2016	1, 3-24	23	11	12	13	10	All
II	2016	101-114, 116-121	20	20	0	12	8	All
III	2016	201-219	19	10	9	14	5	All
IV	2017	501, 524, 525, 534-538, 552-555	12	12	0	12	0	4 ^a
V	2017	502-523	22	22	0	22	0	4 ^a
VI	2017	526-533, 539-551	21	21	0	21	0	4 ^a
			117	96	21	94	23	

^a Hand length; difference 1-3; span 1-5; and span 2-5.

Fig. 2. Examples of the rating of 2 items of the PHE. (A) Span 2-3 and (B) = difference 1-3. PHE = Practical Hand Evaluation.

Potential confounding factors

To evaluate potential influences on the reliability, the data set was stratified on gender, joint hyperlaxity, and item order. For each stratified group within each potential confounding factors, 1-way random ICCs were calculated for the 2 measurements on the first day. Differences in ICCs were visually evaluated to see if a pattern would emerge, and if they fell into identical interpretation classes as defined by Portney and Watkins.²⁹ These interpretation classes were used as a reference to determine whether reliability within a particular group would be different from that within another group. Differences that imply going from the poor to the good or excellent interpretation class were considered unacceptable. To investigate the influence of possible stretch and/or training effects, the differences between the first and second measurements on the first day were assessed using limits of agreement (LoA). The LoA was defined as the mean difference ± 1.96 SD of differences between measurements on days 1 and 2.

Item reduction

Two types of analyses were done to identify collinearity and possibilities for further item reduction. First of all, the variation inflation factor (VIF) with a cutoff of 5 was used for stepwise elimination of items, which could be predicted by other items. In addition, hierarchical clustering of correlation coefficients was used to group correlating items together. Correlation plots for different numbers of clusters were visually inspected for their fit.³¹ Possible item reduction was analyzed using the crude data set as well as data adjusted for hand length.

All data were analyzed using the R software environment, version 3.3.1.³² Level of significance was set at $P \leq .05$, 2-tailed.

Results

About 117 subject participants enrolled in this study; 62 in 2016 and 55 in 2017 (Table 2). No eligible subject participants refused to participate, and no dropouts were identified during the study.

The percentage of missing data per PHE item was acceptable for 11 of 14 items (78.6%), doubtful for 2 of 14 items (14.3%), and not acceptable for 1 item (span 3-4; 7.1%). The latter was caused by the fact that 16 participants were unable to perform the hand position of span 3-4 item.

Reliability

ICCs for intrarater, inter-rater_{short}, and inter-rater_{long} reliability were interpreted as excellent in 14.3%, 14.3%, and 0%, respectively; as good in 78.6%, 85.7%, and 92.9%, respectively; and as moderate in 7.1%, 0%, and 7.1%, respectively; poor reliability was not found (Table 3). Moderate reliability was found for the inter-rater reliability for span 1-5 item and for 3 of 4 values for span 3-4 item. ICCs for intrarater reliability were generally higher than those for the inter-rater reliability. In terms of the classification of Nunally and Bernstein,³³ most ICCs are in the ≥ 0.80 range, indicating that the PHE is reliable for use as a screening instrument to assess anthropometric features and flexibility/stiffness of the hand.

Measurement error

SEM ranged from 2.4 to 8.4 mm for the total sample (Table 4). The percentage of shifts by ≥ 2 categories was between 9.4% and

Table 2
Demographic variables of the study population

Potential confounding factors	Group	Gender		Hyperlaxity		Item order		Total	Unit
		Male	Female	Yes	No	Down	Up		
Body height		182.6 $\pm$ 5.6	170.7 $\pm$ 6.9	173.3 $\pm$ 8.4	176.7 $\pm$ 8.6	176.2 $\pm$ 8.9	174.3 $\pm$ 7.5	175.9 $\pm$ 8.7	Mean cm $\pm$ SD
Age		23.5 $\pm$ 4.5	22.2 $\pm$ 3.2	22.7 $\pm$ 2.4	22.8 $\pm$ 4.3	22.4 $\pm$ 2.9	24.4 $\pm$ 6.5	22.8 $\pm$ 3.9	Mean years $\pm$ SD
Gender	Male	—	—	10 (20%)	40 (80%)	44 (86%)	7 (13%)	51 (100%)	N (group row %)
	Female	—	—	19 (28%)	47 (71%)	52 (79%)	14 (21%)	66 (100%)	
Hyperlaxity	Yes	10 (34%)	19 (66%)	—	—	22 (76%)	7 (24%)	29 (100%)	
	No	40 (46%)	47 (54%)	—	—	73 (84%)	14 (16%)	87 (100%)	
Item order	Down	44 (46%)	52 (54%)	22 (23%)	73 (77%)	—	—	96 (100%)	
	Up	7 (33%)	14 (67%)	7 (33%)	14 (67%)	—	—	21 (100%)	
Total	Total	51 (44%)	66 (56%)	29 (25%)	87 (75%)	96 (82%)	21 (18%)	117 (100%)	

SD = standard deviation.

Table 3
Intrarater and inter-rater reliability

Potential confounding factors	Intrarater ICC (95% CI)	Inter-rater _{short} ICC (95% CI)	Inter-rater _{long} ICC (95% CI)
Hand length	0.90 (0.85–0.93)	0.90 (0.87–0.93)	0.87 (0.82–0.91)
Hand width	0.87 (0.81–0.90)	0.78 (0.69–0.84)	0.80 (0.72–0.85)
Difference 1–3	0.80 (0.73–0.86)	0.83 (0.77–0.88)	0.79 (0.71–0.85)
Difference 3–5	0.77 (0.68–0.83)	0.76 (0.68–0.83)	0.75 (0.65–0.82)
Span 1–2	0.79 (0.71–0.85)	0.75 (0.66–0.82)	0.73 (0.64–0.81)
Span 1–3	0.88 (0.83–0.92)	0.87 (0.82–0.91)	0.80 (0.72–0.85)
Span 1–4	0.87 (0.82–0.91)	0.81 (0.74–0.87)	0.80 (0.72–0.86)
Span 1–5	0.90 (0.85–0.93)	0.90 (0.86–0.93)	0.84 (0.77–0.88)
Span 2–3	0.80 (0.73–0.86)	0.80 (0.73–0.86)	0.79 (0.71–0.85)
Span 2–4	0.85 (0.78–0.89)	0.75 (0.66–0.82)	0.75 (0.66–0.82)
Span 2–5	0.80 (0.72–0.85)	0.86 (0.81–0.90)	0.81 (0.74–0.86)
Span 3–4	0.67 (0.56–0.76)	0.77 (0.69–0.84)	0.76 (0.67–0.83)
Span 3–5	0.78 (0.69–0.84)	0.82 (0.75–0.87)	0.80 (0.73–0.86)
Span 4–5	0.83 (0.76–0.88)	0.80 (0.72–0.86)	0.79 (0.71–0.85)

ICC = intraclass correlation coefficient; 95% CI = 95% confidence interval; Inter-rater_{short} = inter-rater reliability within day 1; Inter-rater_{long} = inter-rater reliability between days 1 and 2.

Bold rows are coefficients based on 117 participants; otherwise, analysis was done on 62 participants.

21.0% in the whole study sample. A shift in the PHE percentiles classification of ≥ 2 categories was found in $<5\%$ for 0 items, in $5\% \leq x < 10\%$ for 4 items (28.6%), in $10\% \leq x < 15\%$ for 10 items (71.4%), and in $\geq 15\%$ for 0 items.

Potential confounding factors

Stratification for gender, joint hyperlaxity, and item order revealed neither major shifts in interpretation classes between categories for these factors nor any significant differences between the 2 rating moments on day 1 (Table 5). Nor did the LoA showed any consistent differences between the raters over all items (range, -2.5 to 1.1 mm), with a random error between 5.7 and 22.0 mm.

The general tendency was that the ICCs among the various stratified groups were all in the same range. Some substantial differences in ICCs were observed (Table 5), for example, a lower reliability for difference 3–5 item in women (which might be related to smaller hand size) and measurements related to the fourth finger in subject participants with hyperlaxity. In combination with a contrary of higher ICCs for other PHE items in case of hyperlaxity, the latter difference might be statistically based.

Table 4
Measurement error properties per item on first test

Item	N	Mean $\pm$ SD (c_v)	Mean _{male} $\pm$ SD (c_v)	Mean _{female} $\pm$ SD (c_v)	SEM (mm)	Percentage _{shift}	% Missing values in all trials
Hand length	117	189.4 $\pm$ 11.9 (6.3%)	198.5 $\pm$ 8.5 (4.3%)	182.5 $\pm$ 9.1 (5.0%)	4.0	9.4	1.3
Hand width	62	82.1 $\pm$ 6.3 (7.7%)	88.1 $\pm$ 4.4 (4.9%)	78.8 $\pm$ 4.6 (5.8%)	2.4	21.0	0.0
Difference 1–3	117	74.6 $\pm$ 7.6 (10.2%)	78.6 $\pm$ 7.2 (9.2%)	71.6 $\pm$ 6.4 (8.9%)	3.4	15.4	1.3
Difference 3–5	62	36.1 $\pm$ 4.1 (11.4%)	36.1 $\pm$ 3.9 (10.8%)	36.0 $\pm$ 4.3 (11.9%)	2.1	17.7	0.0
Span 1–2	62	153.2 $\pm$ 17.1 (11.2%)	159.9 $\pm$ 20.9 (13.1%)	149.5 $\pm$ 13.6 (9.1%)	8.3	16.1	0.0
Span 1–3	62	176.5 $\pm$ 17.4 (9.8%)	186.5 $\pm$ 18 (9.6%)	171.1 $\pm$ 14.5 (8.5%)	6.2	14.5	0.0
Span 1–4	62	181.1 $\pm$ 21.7 (12.0%)	195.4 $\pm$ 22.4 (11.4%)	173.6 $\pm$ 17.3 (9.9%)	8.2	13.8	5.6
Span 1–5	117	194.4 $\pm$ 18.9 (9.7%)	205.9 $\pm$ 18.3 (8.9%)	185.5 $\pm$ 13.9 (7.5%)	6.4	14.5	1.3
Span 2–3	62	74.4 $\pm$ 13.9 (18.7%)	77.1 $\pm$ 15.6 (20.2%)	72.9 $\pm$ 12.8 (17.6%)	6.3	9.7	0.0
Span 2–4	62	96.8 $\pm$ 16.4 (17.0%)	101.2 $\pm$ 15.5 (15.3%)	94.2 $\pm$ 16.6 (17.7%)	7.1	19.6	7.3
Span 2–5	117	134.6 $\pm$ 16.2 (12.0%)	140.6 $\pm$ 16.4 (11.7%)	129.9 $\pm$ 14.4 (11.1%)	7.0	6.8	1.3
Span 3–4	62	58.9 $\pm$ 14.0 (23.8%)	64.8 $\pm$ 13 (20.1%)	55.5 $\pm$ 13.7 (24.7%)	7.5	6.8	25.8
Span 3–5	62	108.8 $\pm$ 14.1 (12.9%)	114.2 $\pm$ 14.4 (12.6%)	105.9 $\pm$ 13.2 (12.4%)	6.3	11.3	0.0
Span 4–5	62	71.5 $\pm$ 14.0 (19.6%)	76.0 $\pm$ 14.4 (18.9%)	68.9 $\pm$ 13.3 (19.3%)	5.9	13.3	2.0

SD = standard deviation; c_v = coefficient of variance; SEM = standard error of measurement.

Bold rows are based on 117 participants; otherwise, analysis was done on 62 participants.

Item reduction

Stepwise backward elimination of items based on the VIF revealed multiple redundant items in the crude and adjusted data set (Table 6). Three of these items concerned measuring the span between the thumb and other digits. This might indicate that all thumb span items provide similar information about the flexibility of the thumb.

Hierarchical clustering of correlation coefficients using various numbers of clusters was visually inspected. Six clusters digit (dig. 1 cluster, dig. 5 cluster, dig. 2–4 cluster, hand size cluster, difference 3–5 cluster, and difference 1–3 cluster) seemed to best fit the correlation plot in the crude and adjusted data sets (Fig. 3). Three of these clusters seemed related to hand/finger size. Three other clusters showed high correlations with span items containing a similar joint. Items related to the thumb and little finger particularly stood out.

Discussion

This study demonstrated good reliability for nearly all PHE items. Measurement errors, however, were substantial relative to variances in the reference population. Variables that were assumed to be of potential influence on the reliability of the PHE (gender, joint laxity, and order of administration) showed no substantial influence. Items could be statistically clustered into 6 separate clusters. The clusters thus identified seem to be physiologically plausible with respect to hand function. Our findings suggest that the PHE meets many of the criteria for screening. The clinical applicability may be increased in the future by using a short version of the PHE, based on our proposed item reduction.

Reliability

This is the first study to present reliability parameters for a practical screening of anthropometric features of the hand. Previously, the test-retest reliability (with 1 rater) and measurement error of hand radiographs were found to be excellent (ICCs, 0.99).⁵ However, the radiographic method has disadvantages of cost, time, radiation load, and the availability of the X-ray equipment. These ICCs should be viewed critically as clinical practice has to deal with different X-rays interpreters, thus dealing with 2 sources of systematic errors. Moreover, with the X-ray method, only anthropometric values were determined, whereas the PHE measures a combination of anthropometric aspects and hand flexibility.

Table 5Inter-rater_{short} ICC of data stratified by gender, hyperlaxity, and item order

Item	Inter-rater _{short} ICC						LoA ^a
	Gender		Hyperlaxity		Item order		
	Male	Female	Yes	No	Up	Down	
Hand length	0.82	0.83	0.91	0.90	0.87	0.91	−1.16 ± 10.83
Hand width	0.53	0.51	0.58	0.85	0.87	0.73	−0.31 ± 6.70
Difference 1-3	0.80	0.77	0.86	0.83	0.76	0.84	1.13 ± 8.96
Difference 3-5	0.59	0.84	0.67	0.80	0.68	0.80	−0.15 ± 5.72
Span 1-2	0.77	0.70	0.91	0.65	0.81	0.74	−1.50 ± 21.66
Span 1-3	0.88	0.80	0.95	0.81	0.83	0.89	−0.55 ± 16.44
Span 1-4	0.88	0.65	0.84	0.78	0.73	0.84	−0.77 ± 22.00
Span 1-5	0.87	0.85	0.94	0.88	0.89	0.90	−0.79 ± 17.63
Span 2-3	0.78	0.83	0.75	0.83	0.86	0.78	−2.47 ± 15.91
Span 2-4	0.76	0.73	0.48	0.83	0.74	0.76	−2.07 ± 18.86
Span 2-5	0.83	0.87	0.88	0.86	0.87	0.86	−1.35 ± 18.93
Span 3-4	0.75	0.76	0.47	0.87	0.83	0.75	−1.07 ± 19.48
Span 3-5	0.75	0.86	0.87	0.80	0.78	0.85	−0.55 ± 16.85
Span 4-5	0.91	0.70	0.73	0.83	0.78	0.82	−1.87 ± 16.72

ICC = intraclass correlation coefficient; LoA = limits of agreement.

Limits of agreement between the first and second measurements on the first day.

Bold rows have coefficients based on 117 participants; otherwise, analysis was done on 62 participants.

^a Paired sample *t* test between scores during first and second measurements on the first day were not significant for all items.

Measurement error

It remains unclear from our study what the clinical relevance of the measurement error might be. We assume that only extreme values of a particular PHE item in a manual worker individual might correlate with MSHC: extremely small hands might be easily overstretched, and extremely large hands might be positioned less efficiently or accurately. Hence, the risk of misclassification of a particular measurement toward the extremes (due to SEM), whereas the actual value is in the midrange of the reference values, should be minimal. The same applies to the shift from an extreme to a midrange category. Thus, although the SEM might be regarded as relatively small, the percentage of shifts (taking the measured value plus or minus SEM into account) relative to reference categories for a larger sample of participants¹⁴ appears to be substantial. On the other hand, Wagner³⁴ indicated that the items of the PHE together form a hand profile of features, and it is not clear whether a harmonic hand profile (ie, the measurement values of 1 hand are within the same or adjoining category) can shift to a disharmonic hand profile (ie, with some values in the upper extreme and others in the lower extreme) due to measurement error. A substantial percentage of shifts implies a lower

specificity of the PHE, with a possibly substantial number of false-positive or false-negative values due to measurement error. This is not an a priori argument against using the PHE as a screening instrument in hand therapy, but in view of the interindividual variability between measurements, care should be taken when interpreting results on an individual basis. The combination of high ICCs and substantial SEM might be interpreted as arguments against the use of the PHE as a screening instrument for hand features. However, the ICCs indicate that the instrument's reliability is sufficient to use it for screening.³³ Moreover, the SEM is a statistical criterion for the maximum difference among measurements at group level, with a 95% certainty that the difference found is a true difference due to differences between the test and retest situation. The percentage of shifts in classification found in this study might therefore be interpreted as a conservative interpretation of the instrument's reliability. The interpretation of a combination of high ICCs and substantial SEM is common in human research. In detailed testing, as is the case with the PHE, a substantial SEM cannot be ignored for screening at the level of the individual.³⁵ A way to reduce SEM is repeated measurement, for example, as shown in repeated vs mono blood pressure measurement.³⁶

Table 6

Data of stepwise backward elimination of item based on collinearity (VIF score)

Crude (not adjusted for hand length)					Item	Adjusted for hand length			
Step 1	Step 2	Step 3	Step 4	Step 5		Step 4	Step 3	Step 2	Step 1
5454	5398	4871	4641	4547	Hand length	—	—	—	—
4228	4178	4168	3508	3417	Hand width	1695	1830	2042	2075
2565	2554	2381	2278	2221	Difference 1-3	1675	1684	1868	1882
2066	1970	1904	1904	1904	Difference 3-5	1325	1346	1365	1401
6944	6568	6334	1990	1937	Span 1-2	1607	4622	5937	6210
17,870	14,076	7507			Span 1-3			13,056	16,274
25,372					Span 1-4				18,832
20,847	15,595				Span 1-5		5671	10,724	13,884
	4676	4514	4470	3816	Span 2-3	3402	3425	3592	3639
5427	4215	4207	4201	4137	Span 2-4	3039	3048	3049	3706
9872	7668	7269	6938		Span 2-5	4024	4486	4487	5800
3075	2490	2462	2337	2261	Span 3-4	2056	2066	2184	2699
4351	4351	3972	3971	2950	Span 3-5	2157	2298	2373	2382
4779	3756	3629	3436	2676	Span 4-5	2365	2370	2561	3372

VIF = variation inflation factor.

Bold rows have variation inflation factors based on 117 participants; otherwise, analysis was done on 62 participants.

Fig. 3. (A) Clustering of the PHE items based on correlation coefficients. (B) Clustering after adjustment for hand length. PHE = Practical Hand Evaluation.

Potential confounding factors

In contrast to our initial assumptions, we did not find any substantial influence of the factors of gender, hyperlaxity, and order of administration. Because the sample data in the studies by Wagner^{14,34} were categorized according to gender, we wanted to know if reliability differed between men and women. The gender factor consisted of 2 characteristics, viz, hand size and percentage of hyperlaxity, which might be relevant in this respect. Women generally have smaller hands than men, and among women, there is a higher percentage of hyperlaxity (both in the general population and in our study sample). The possible influence of differences in hand size is discussed previously. The presence of hyperlaxity showed no substantial influence on reliability. Together with our finding that the reliability of the PHE is not influenced by the order of item administration (due to stretching effects), the absence of an influence of these factors simplifies the test instructions for future raters. This is another advantage of the PHE when used in screening situations. No statistical differences were found between the first and second measurements on the first day, suggesting limited effects of training and stretching on the second measurement.

Item reduction

Although a number of subjects failed to meet the requirements for a full analysis on possible item reduction, an exploratory analysis was done to evaluate clustering among PHE items. Stepwise elimination of items based on the VIF showed that span items that included the thumb had high collinearity. What these scores have in common is thumb length and flexibility of the thumb joint. This might indicate that these components make major contributions to the span items. All thumb-related items might thus be replaced by a single item span 1-2 or span 1-5, depending on clinical task relevance. Hierarchical clustering of correlation coefficients of items supported this hypothesis. In addition to dig. 1 cluster, other clusters with assumed clinical relevance were also identified. Three of them were related to span and 3 others to hand/finger size. Scores adjusted for hand length showed more distinctive clustering compared with crude scores, so hand length might be an

underlying confounding factor that needs to be corrected for in future use. Similarly, no clinical interpretation is as yet available of individual or combined scores in terms of specific symptoms. Wagner^{14,34} suggested that some individual items might be associated with specific pain features, for instance, pianists with span 1-5, which is too small to reach a full octave on a regular piano. But disharmony between multiple item percentile scores could also be an indicator of hand feature deficiencies. In conclusion, our preliminary analysis yielded some starting points for reducing the number of items in the PHE even further. Both statistical clustering and clinical predictability should be kept in mind when restructuring items of the PHE for future research.

Strengths and limitations

The design of our study was chosen so as to maximize the reliability of the study.³¹ In theory, reliability measures may display variation arising from several sources; the number of participants and number of measurements, the ranges between the minimum and maximum values of the data, the measurement instrument itself, the raters, the participants, the circumstances under which the measurements are made, and random factors.²⁶ Three of these aspects are discussed successively later. One of the strengths of this study was the large number of participants. However, data for 10 of 14 PHE items were based on only 62 of the 117 participants; if it had been based on more participants, even higher ICCs for these items might have been found. The range from the minimum to maximum number of millimeters per item differed substantially. It can be assumed that for the smallest distances (eg, difference dig. 3-5), the impact of the SEM could be bigger than for larger distances (eg, difference dig. 1-5) as the rater cannot reduce the measurement error below a certain number of millimeters (eg, due to minor changes in the position of the pen while marking the distance on paper). Thus, the potential impact of the SEM on the reliability of the 2 items with the smallest mean distance (ie, differences dig. 3-5 and dig. 2-3) can be assumed to be bigger than its impact for other PHE items. The potential impact of the SEM on the reliability of most items was substantial, with a slight tendency toward lower ICCs and higher percentages of shifts

for the 2 aforementioned items with small mean distances. The subject participants were healthy Dutch volunteers and aged between 17 and 50 years. Although this had the advantage of a relatively homogeneous study sample, the generalizability of the findings toward other populations might be limited to some degree. Our population was taller than the reference population from Germany.¹⁴ As smaller hands may increase the impact of the shift percentage due to SEM, this may reduce the clinical applicability. Because our study excluded patients with MSHC, it is not clear to what extent the reliability of the PHE may be generalized to the population of patients with MSHC. The raters were extensively instructed, and we assume that this instruction has greatly benefited the reliability of the study.

Clinical implications

A systematic approach to the examination of the hand increases the chance of detecting anthropometric hand features that might relevantly interfere with the desired high level of hand functioning^{14–16} or that might pose a risk for the development of MSC. The PHE can be used for research purposes at group level, in view of its high ICCs. The PHE can also be used for individual screening of hand profiles, despite the substantial SEM, as the 14 PHE items together contain redundant information (as shown by our VIF analysis). Thus, the impact of the substantial SEM will be reduced in an individual screening if all 14 items are assessed together. At this stage of the development of the PHE, the individual items of the PHE alone should not be used for the assessment of individual anthropometric hand features. Reducing the SEM will be needed to overcome this shortcoming.

Future research

As stated previously, in future studies, the reliability of the PHE might be improved by having each rater measure each item 2 or 3 times. Further research might consider determining the reliability of a digital photographic alternative for the manually measured PHE procedure. In combination with a reduction in the number of PHE items, such a photographic PHE procedure could meet more of the criteria for a screening instrument because of fewer false-positive or false-negative findings, even if repeated photographs are needed. Another suggestion for further research might be a prospective study of the correlation of the difference between harmonic and disharmonic hand profiles and the prevalence of MSC of the hand. Although a causal relation seems plausible, there is so far no evidence about the so-called at risk hand profiles for MSC of the hand when combined to specific work demands.

Conclusion

Results of this study showed that the observed agreement in test-retest administration of the PHE within and between raters was high for most PHE items. Among the items, the reliability of measuring span 3–4 showed the lowest agreement. Measurement errors were substantial relative to variances found in a reference population. Our results, however, suggest that the PHE is a suitable instrument for the screening of anthropometric hand features. Gender, joint hyperlaxity, and order of administration did not substantially influence its reliability. Statistical clustering of items was possible for several items of the PHE, with 6 clusters that seem to be physiologically plausible with respect to hand function. Together, these results suggest that the PHE meets many of the

criteria for screening. Clinical value might be further improved by item reduction.

References

1. Arslan Y, Bülbül I, Öcek L, Şener U, Zorlu Y. Effect of hand volume and other anthropometric measurements on carpal tunnel syndrome. *Neurol Sci.* 2017;38(4):605–610.
2. Mondelli M, Curti S, Farioli A, et al. Anthropometric measures as a screening test for carpal tunnel syndrome: receiver operating characteristic curves and accuracy. *Arthritis Care Res.* 2015;67(5):691–700.
3. Mondelli M, Curti S, Mattioli S, et al. Associations between body anthropometric measures and severity of carpal tunnel syndrome. *Arch Phys Med Rehabil.* 2016;97(9):1456–1464.
4. Lim PG, Tan T, Ahmad S. The role of wrist anthropometric measurement in idiopathic carpal tunnel syndrome. *J Hand Surg Eur.* 2008;33(5):645–647.
5. Paul SN, Kato BS, Hunkin JL, Vivekanandan S, Spector TD. The big finger; the second to fourth digit ratio is a predictor of sporting ability in women. *Br J Sports Med.* 2006;40(12):981–983.
6. Phelps VR. Relative index finger length as a sex influenced trait in man. *Am J Hum Genet.* 1952;4(2):72–89.
7. Honekopp J, Manning T, Muller C. Digit ratio (2D:4D) and physical fitness in males and females: evidence for effects of prenatal androgens on sexually selected traits. *Horm Behav.* 2006;49(4):545–549.
8. Kok LM, Huisstede BM, Voorn VM, Schoones JW, Nelissen RG. The occurrence of musculoskeletal complaints among professional musicians: a systematic review. *Int Arch Occup Environ Health.* 2016;89(3):373–396.
9. Silva AG, Afreixo V. Pain prevalence in instrumental musicians: a systematic review. *Med Probl Perform Art.* 2015;30(1):8–19.
10. Zaza C, Charles C, Muszynski A. The meaning of playing-related musculoskeletal disorders to classical musicians. *Soc Sci Med.* 1998;47(12):2013–2023.
11. Pascarelli EF, Hsu YP. Understanding work-related upper extremity disorders: clinical findings in 485 computer users, musicians and others. *Eur J Appl Physiol Occup Physiol.* 1999;79(2):127–140.
12. Baadjou VAE, Roussel NA, Verbunt JMF, Smeets RJE, de Bie RA. Systematic review: risk factors for musculoskeletal disorders in musicians. *Occup Med (Lond).* 2016;66(8):614–622.
13. Münte TF, Altenmüller E, Jäncke L. The musician's brain as a model of neuroplasticity. *Nat Rev Neurosci.* 2002;3(6):473–478.
14. Wagner C. *Hand und Instrument. Musikphysiologische Grundlagen, Praktische Konsequenzen.* Wiesbaden, Germany: Breitkopf & Härtel; 2005.
15. Wilson FR, Wagner C, Hömberg V. Biomechanical abnormalities in musicians with occupational cramp/focal dystonia. *J Hand Ther.* 1993;6:298–307.
16. Leijnse JNAL, Bonte JE, Landsmeer JMF, Kalker JJ, van der Meulen JC, Snijders CJ. Biomechanics of the finger with anatomical restrictions—the significance for the exercising hand of the musician. *J Biomech.* 1992;25:1253–1264.
17. Faria J, Ordóñez FJ, Rosity-Rodríguez M, et al. Anthropometrical analysis of the hand as a repetitive strain injury (RSI) predictive method in pianists. *Ital J Anat Embryol.* 2002;107(4):225–231.
18. Sakai N. Hand pain attributed to overuse among professional pianists: a study of 200 pianists. *Med Probl Perform Art.* 2002;17(4):178–180.
19. Wristen BG, Jung MC, Wisner AKG, Hallbeck MS. Assessment of muscle activity and joint angles in small-handed pianists: a pilot study on the 7/8-sized keyboard versus the full-sized keyboard. *Med Probl Perform Art.* 2006;21(1):3–9.
20. Yoshimura E, Mia Paul P, Aerts C, Chesky K. Risk factors for piano-related pain among college students. *Med Probl Perform Art.* 2006;21(3):118–125.
21. Chesky K, Yoshimura E. Hand size and PRMDs in Japanese female pianists. Letter to the Editor. *Med Probl Perform Art.* 2007;22:39–40.
22. Linari-Melfi M, Cantarero-Villanueva I, Fernández-de-las-Peñas C, Guisado-Barillao R, Arroyo-Moralis M. Analysis of deep tissue hypersensitivity to pressure pain in professional pianists with insidious mechanical neck pain. *BMC Musculoskelet Disord.* 2011;12:268.
23. Bragge P, Bialocerkowski A, McMeeken J. A systematic review of prevalence and risk factors associated with playing-related musculoskeletal disorders in pianists. *Occup Med (Lond).* 2006;56(1):28–38.
24. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol.* 2011;64(1):96–106.
25. Grahame R, Bird HA, Child A. The revised (Brighton 1998) criteria for the diagnosis of benign joint hypermobility syndrome (BJHS). *J Rheumatol.* 2000;27(7):1777–1779.
26. Vet de HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine.* Cambridge, GB, England: Cambridge University Press; 2011.
27. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods.* 1996;1(1):30–46.
28. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15(2):155–163.
29. Portney LG, Watkins MP. *Part IV Data Analysis: Correlation. Foundations of Clinical Research. Applications to Practice.* 2nd ed. Upper Saddle River, NJ: Prentice Hall Health; 2002.
30. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The standard error of measurement is a more appropriate measure of quality for postgraduate

medical assessments than is reliability: an analysis of MRCP (UK) examinations. *BMC Med Educ.* 2010;10:40.

31. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013. <http://www.R-project.org/>.
32. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. *Qual Quant.* 2007;41(5):673.
33. Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd ed. Hillsdale, NJ: McGraw-Hill; 1994.
34. Wagner C. Musicians' hand problems: looking at individuality. A review of points of departure. *Med Probl Perform Art.* 2012;27(2):57–64.
35. Denegar CR, Ball DW. Assessing reliability and precision of measurement: an introduction to intraclass correlation and standard error of measurement. *J Sport Rehab.* 1993;2:35–42.
36. Yong F, Heiss G, Couper D, Meyer ML, Cheng S, Tanaka H. Measurement repeatability of central and peripheral blood pressures: the ARIC study. *Am J Hypertens.* 2017;30(10):978–984.