Patient-centered outcomes in hand surgery

Ghady El Khoury, MD

Thesis submitted to obtain the degree of

Doctor in Medical Sciences

Promoters

Professor Philippe Lefèvre

Professor Olivier Barbier

November 2021

A Cléore, Olivia et Charline

A Anaïs

A mes parents

"C'est le temps que tu as perdu pour ta rose qui fait ta rose si importante"

Antoine de Saint-Exupéry

Remerciements

Au moment d'écrire ces lignes, je me rends compte que quatre années se sont écoulées depuis le début de cette aventure. Une petite pause dans mon parcours de clinicien, néanmoins enrichissante et épanouissante. Je tiens à remercier toutes les personnes qui étaient à mes côtés et ont rendu ce parcours assez agréable.

Merci à tous les membres de mon comité d'accompagnement qui se sont donnés à fond pour l'aboutissement de mon travail. Merci à mes promoteurs les professeurs Philippe Lefèvre et Olivier Barbier pour leur confiance, leur esprit critique et leur positivité. Merci Xavier pour ton amitié, Massimo pour ton implication inconditionnelle et ta rigueur, Jean-Louis pour ton soutien et ton attitude toujours positive. Merci également aux membres externes du jury les professeurs Esther Vögelin et Ingemar Merkies pour leurs commentaires encourageants.

Merci à tous les membres du laboratoire COSY de l'Institut de Neurosciences avec qui j'ai pu partager de bons moments, malgré le télétravail. Merci à mes collègues de bureau, Laurent et Félicien, pour leur bonne humeur et leur aide précieuse. Merci à tous mes collègues assistants en orthopédie qui ont facilité le recrutement des patients.

Merci également à toute ma famille libanaise, auprès de laquelle il est toujours bon de se ressourcer. Merci à mes parents, pour ce que je suis, pour leur amour inconditionnel et l'ambition qu'ils m'ont inculquée. Merci à mon frère Pascal qui n'hésitait pas à me proposer ses compétences de codeur. Merci à ma famille belge, pour le soutien, les encouragements et les bulles (assez fréquentes) de champagne.

Merci à Cléore, Olivia et Charline mes petits rayons de soleil qui ont rajouté du challenge et de la joie à mon quotidien. Merci à Anaïs, ma partenaire dans les moments les plus décisifs de cette vie. Merci pour ton amour, tes idées brillantes et tes conseils précieux.

Members of the jury

President:	**Prof. André Mouraux** Institute of Neuroscience, Université catholique de Louvain, Belgium
Promoters :	**Prof. Philippe Lefèvre** Institute of Information and Communication Technologies, Electronics and Applied Mathematics, Université catholique de Louvain, Belgium
	Prof. Olivier Barbier Institut de recherche expérimentale et clinique, Université catholique de Louvain, Belgium
Members :	**Prof. Jean-Louis Thonnard** Institute of Neuroscience, Université catholique de Louvain, Belgium
	Prof. Xavier Libouton Institut de recherche expérimentale et clinique, Université catholique de Louvain, Belgium
	Dr. Massimo Penta Institute of Neuroscience, Université catholique de Louvain, Belgium
External members:	**Prof. Esther Vögelin** Inselspital, University Hospital of Bern, Switzerland
	Prof. Ingemar Merkies Maastrich University Medical Center, the Netherlands

Contents

CHAPTER 1

Introduction

In hand surgery, given the variety of available surgical techniques, the choice of the "right" treatment (the treatment that gives the best results) is complicated, even for experienced surgeons. Nowadays, the surgeon's opinion is no longer sufficient; a detailed clinical exam must include objective assessments as well as the patient's perception so that the post-operative outcome can match the patient's expectations. Hand surgery practice is shifting towards evidence-based treatments with the aim of providing the best results when treating patients. Therefore, health professionals need robust tools to evaluate objectively the effectiveness of a surgical treatment of the hand. This thesis explores new tools the hand surgeon can use for the evaluation of treatment effects.

1.1. Hand pathologies and function impairment

The hand is a highly specialized functional, sensory, and aesthetic unit. It can suffer a unique and wide range of lesions such as bone fractures, nerve compressions, ligament and tendon injuries, cartilage degeneration, vascular lesions and skin conditions. In addition, hand functioning can be impaired by pathologies that affect the brain-hand connection such as stroke or neuro-degenerative diseases. Hand injuries and pathologies can cause functional loss in young and active workers (e.g. wrist fracture when skiing), as well as in the older population (e.g. osteoarthritis). Irrespective of the etiology, impaired hand function limits the individual's ability to perform activities of daily living (ADL), and increases demands on caregivers and workers' compensation costs (Desrosiers et al., 2006; Kwakkel et al., 2003). When the costs of medical care, rehabilitation, and productivity loss are computed, the burden of hand disorders is massive (de Putter et al., 2012). Therefore, the evaluation and management of hand disorders is critical to individuals and to society. As health care delivery and reimbursement undergo rapid and substantial changes, the focus on quality and value of care continues to increase (Giladi and Chung, 2013), with the aim of reducing the costs of medical care, and increasing the use of high-value treatments, while

discouraging low-value treatments. This is done by emphasizing appropriate and evidence-based surgical interventions or conservative treatments.

The first step in determining the appropriate treatment for a given patient is to accurately identify and evaluate the impaired function. Based on an accurate and reliable assessment, clinicians are able to establish treatment goals. Therefore, as a complement to the clinical examination and medical history, many tests have been developed such as radiography to visualize bones, goniometry to measure range of motion, or dynamometry to measure grip strength. However, the relation between impairments (dysfunction at the organ/segment level) and disability (dysfunction at the person level) is not straightforward (Arnould et al., 2007; Vandervelde et al., 2009). This means that one cannot deduce the ability to carry out activities of daily life just by looking at the impaired functions (such as loss in grip strength or range of motion). For this reason, patient functioning in daily life must be measured per se, and not merely inferred from underlying impairments. Therefore, clinicians need to use outcome measures (i.e. assessments) that target the adequate domain and have strong psychometric properties, as will be developed in the next chapters.

1.2. Outcome measurement

Treatment methods have benefits, as well as associated risks, disadvantages and costs. Therefore, robust criteria are needed to justify these costs and to reinforce patient education regarding the risks and outcomes of a procedure. Outcome evaluations are needed to assess the effectiveness and reliability of a treatment. The trend in the reimbursement of treatments by private or social insurances is shifting from "pay for an act" to "pay for results" (Porter, 2009), with the goal of achieving high value for patients (Porter, 2010).

Physical tests and biomechanical measurements (such as grip strength, pinch strength, or range of motion) are the most commonly reported outcome measures in hand surgery (Alderman and Chung, 2008). These simple empirical measures of physical function are easily captured by clinicians and therapists, and are helpful metrics of functional recovery after hand surgery (Klum et al., 2012). Although objective and reproducible measurements can be obtained, these outcomes may not be the most appropriate to reflect the true benefit of a treatment to the patient. Indeed, these methods do not reflect the patient's ability to carry out activities of daily living, the ability to return to previous occupations, and pain. What the surgeon might view as a considerable improvement in grip strength may not correspond to improved hand function from the patient's perspective (Giladi and Chung, 2013). Likewise, demonstration of fracture union on a radiograph is insufficient to determine whether a patient is satisfied with their outcome and is capable of resuming their usual activities or returning to work (Giladi and Chung, 2013; Jaremko et al., 2007; Synn et al., 2009; Young and Rayan, 2000).

In hand surgery, most of the procedures aim at improving functional outcomes and giving patients better quality of personal and social life (Dubert, 2014). Evaluating such outcomes requires patient-centered instruments, as patients are in the best position to describe their stated of health, functional capacity and satisfaction. Accordingly, Patient-Reported Outcome Measures (PROMs) have been developed. PROMs were once considered subjective and unreliable, but are now recognized as fundamental to understanding the impact of clinical decisions (MacDermid, 2014). Contrary to the sole use of clinical evaluation, the inclusion of outcome measures completed by patients themselves is based on the principle that no one is better placed than patients to know their own needs and criteria for results. Patients' self-evaluation is especially adapted to certain aspects of health such as daily activities, satisfaction, social well-being, pain and quality of life (Dubert, 2014).

Since their introduction, the quality of PROMs has improved and their performance is evaluated using criteria similar to clinical measurement tools, such as validity, reliability, and sensitivity. It has become increasingly recognized that new drugs, devices and interventions must prove themselves in terms of better outcomes at the patient level to warrant private or public investments. For example, the Belgian Federal Institute for Health Insurance (INAMI, institut national d'assurance maladie-invalidité) has adopted PROMs and cost-efficiency as outcome measures for the evaluation of a mobile application for the rehabilitation after hip and knee replacements (INAMI, 2021). On a global scale, the International Consortium for Health Outcomes Measurement (ICHOM) has been set up as a non-profit organisation with the purpose of transforming health care systems worldwide by measuring and reporting outcomes that matter most to patients (ICHOM, 2021a). In this respect, PROMs constitute a substantial part of the standard evaluation sets developed by ICHOM (ICHOM, 2021b). The research community has also recognized the importance of PROMs, and most large clinical trials now use them as the primary outcome of interest to determine the effectiveness of interventions, with impairment and imaging considered as secondary measures (Chen et al., 2021).

1.2.1. Patient-reported outcome measures in hand surgery

A wide variety of PROMs have been developed for the evaluation of upper extremity disorders, including those for the evaluation of wrist and hand function. These questionnaires are classified as generic, system-specific and disease-specific (Fitzpatrick et al., 1998).

Generic instruments are intended to capture a very broad range of aspects of health status, without focusing on any specific disease or organ system. The Short Form (SF)-36 (Ware et al., 1994), and its shortened version the SF-12 (Ware et al., 1996) are generic measures frequently used as an outcome measure in hand surgery. They ascertain general well-being,

including components of pain, vitality, emotional and mental health, and self-assessment of ability to perform daily functions and activities. The main advantage is that this type of instrument can be used for a broad range of conditions, thereby allowing comparisons of health outcomes across different pathologies and fields. By including items across a broad range of aspects of patients' life, generic instruments must sacrifice some level of detail in terms of relevance to any one illness. The risk is therefore some loss of relevance of questionnaire items when applied to any specific context, and the loss of sensitivity to change that might occur as a result of an intervention.

System-specific instruments focus on an organ system or functional unit. They assess health problems in a specific part of the body. The most commonly used instruments of this type in upper extremity studies are the Michigan Hand Outcomes Questionnaire (MHQ) (Chung et al., 1998), the Disabilities of the Arm, Shoulder and Hand (DASH) (Hudak et al., 1996), and the Patient-Rated Wrist Evaluation (PRWE) (MacDermid et al., 1998). The MHQ assesses each hand independently, and provides data for overall hand performance as well as unique scores of separate domains related to hand function, daily activities, work performance, pain, aesthetics, and satisfaction. The DASH is a self-administered questionnaire of 38 items designed to measure disability for any region in the upper limb. The subjects are asked to rate their ability to carry out activities of daily life regardless of the limb needed to perform that activity. As such, the questionnaire produces a score of patient function representing the composite abilities of both upper extremities. The QuickDASH (Beaton et al., 2005) is the 11-item shortened version of the DASH that was developed to minimize time and responder burden. The PRWE is a 15-item questionnaire measuring wrist pain and function during daily activities. The PRWE stem questions have been modified to allow its application for wrist and hand problems (MacDermid and Tottenham, 2004).

Disease-specific instruments focus on a particular disorder. Their focused nature often results in high responsiveness when used in the appropriate patient population (Szabo, 2001). One of the drawbacks is that the design of such questionnaires often limits their use for the evaluation of other diseases. For instance, the Boston Carpal Tunnel Questionnaire (CTQ) evaluates the severity of symptoms and functional status associated with carpal tunnel syndrome (Levine et al., 1993).

1.2.2. Manual activities monitoring

Ambulatory monitoring devices are enabling a new paradigm of health care by collecting and analyzing data for reliable diagnostics or patient follow-up. This is the case in many fields of medicine, especially cardiology (Sana Furrukh et al., 2020). Monitoring devices are also increasingly present in our daily lives in the form of wearable instruments such as sports watches. These could offer the opportunity to collect data from everyday life for healthcare purposes such as diagnostic, evaluation, and rehabilitation purposes.

A great number of patients have activity limitations caused by impairment of the upper extremities. PROMs in the form of questionnaires have been developed to assess the disability and recovery of the upper limb, as previously seen. These types of tests are very useful to gather information about patients' ability to perform their daily life activities, through self-perceived performance. However, these tests do not generate information about the number of daily activities actually performed by patients in their natural environment. Additionally, many patients can overcome activity limitations by executing activities in a different manner, such as using two hands for an activity that usually requires one (Barbier et al., 2003). Common examples of activities for which patients can develop compensatory mechanisms are typically unimanual activities such as brushing teeth, writing and drinking. Therefore, measuring the actual amount of daily

activity performed by patients is essential to understand the impact of their restrictions on their daily lives. An objective evaluation of activities could complement the input of patients through questionnaires.

An instrument that measures 1) the activities that are actually performed with the hands, 2) the quantity of activity execution (i.e. the amount of hand use), and 3) the quality of execution of the movement, is currently lacking. A device that could be used for patients monitoring in their natural environment should meet the following specifications: 1) the measurement should be objective, i.e. not requiring subjective interpretations by the patient or clinician, 2) the instrument should be portable and unobtrusive for ambulatory use in daily life conditions, 3) the instrument should be able to identify specific activities and provide measures of the quality of activity performance, and, finally 4) the instrument should be applicable in different patient populations (Lemmens et al., 2015).

Wearable inertial sensors are the most common devices for the measurement of motion and physical activities associated with daily living. They combine an accelerometer, a gyroscopic sensor, and sometimes a magnetometer, which makes them particularly effective for evaluating movements (Tamura, 2014). Accelerometers measure acceleration, gyroscopes measure angular velocity, and magnetometers measure magnetic fields (i.e. the orientation towards the Earth's magnetic field). Inertial sensors have been used for monitoring activities as they are small, affordable, and generally unobtrusive (Yang and Hsu, 2010). They have been used for upper limb motion analysis with good accuracy and reliability (Cuesta-Vargas et al., 2010; Zhou et al., 2008). They have been shown useful for clinical applications (Thanawattano et al., 2015), and proved to be more sensitive than questionnaires to detect changes in shoulder movement, thus adding a complementary objective component to outcome measurement (Körver et al., 2014).

The ability to monitor activities of daily living in the patient's natural environment could thus become a valuable tool for clinical decision-making, evaluating healthcare interventions and tracking rehabilitation progress. The process of developing a manual activities monitoring device will be discussed in chapter 5.

1.3. Theoretical framework of measurement

Health status is a multi-faceted concept. Therefore, its accurate and uncontroversial measurement is complex and elusive (Ziebland et al., 1993). Improvements observed by the clinician may not necessarily correspond to the patient's perceptions and experiences, as patients are the best positioned to judge their levels of disability and health-related quality of life (Berkanovic et al., 1995; Hewlett, 2003). The development of PROMs, has been a development of quite revolutionary significance as it allowed to capture the multiple facets of health. Respondents are asked to report on their ability to perform tasks, their energy and sleep patterns, their mood state, experience of pain, social activities, physical mobility and dexterity. The inclusion of such diverse domains comes with the burden of accurately capturing these aspects and measuring them reliably. In this section, the theoretical framework for measuring health outcomes will be exposed, as well as measurement theories for quantifying such variables.

1.3.1. The ICF classification

When choosing an instrument to assess an outcome in clinical practice or research, it is important to consider the construct or domain to be measured, and to evaluate its appropriateness in the given context. PROMs, like any other measurement tool, have specific measurement properties in terms of their scope. It is particularly useful to understand and measure health outcomes that look beyond mortality and morbidity, as they reflect a

biopsychosocial perspective describing the impact of a disease from an individual and societal perspective (Jerosch-Herold et al., 2006).

The adoption of the World Health Organization's International Classification of Functioning, Disability and Health (ICF) has changed the way health and disability are viewed (World Health Organization, 2001). According to the ICF, the consequences of a disease are considered in three domains: (1) body functions and structures, (2) activity, and (3) participation. Body functions refer to physiological and psychological function of the body systems (e.g. motor skills or sensitivity), and body structures are anatomical parts of the body such as bones, muscles and ligaments. Activity is defined as the execution of a task or action by an individual (e.g. manual activities of daily living). Participation refers to the patient's involvement in society, such as in hobbies and work. The impact of a pathology or a surgical intervention in these three domains is also conditioned by personal factors (e.g. motivation, capacity to develop compensatory strategies) and environmental factors (e.g. social or professional context). Pathologies can affect each domain and a patient can have (1) impairments, (2) activity limitations, and (3) participation restrictions. The ICF provides a framework for classifying diseases and their effect on body structure and functioning, activities and participation.

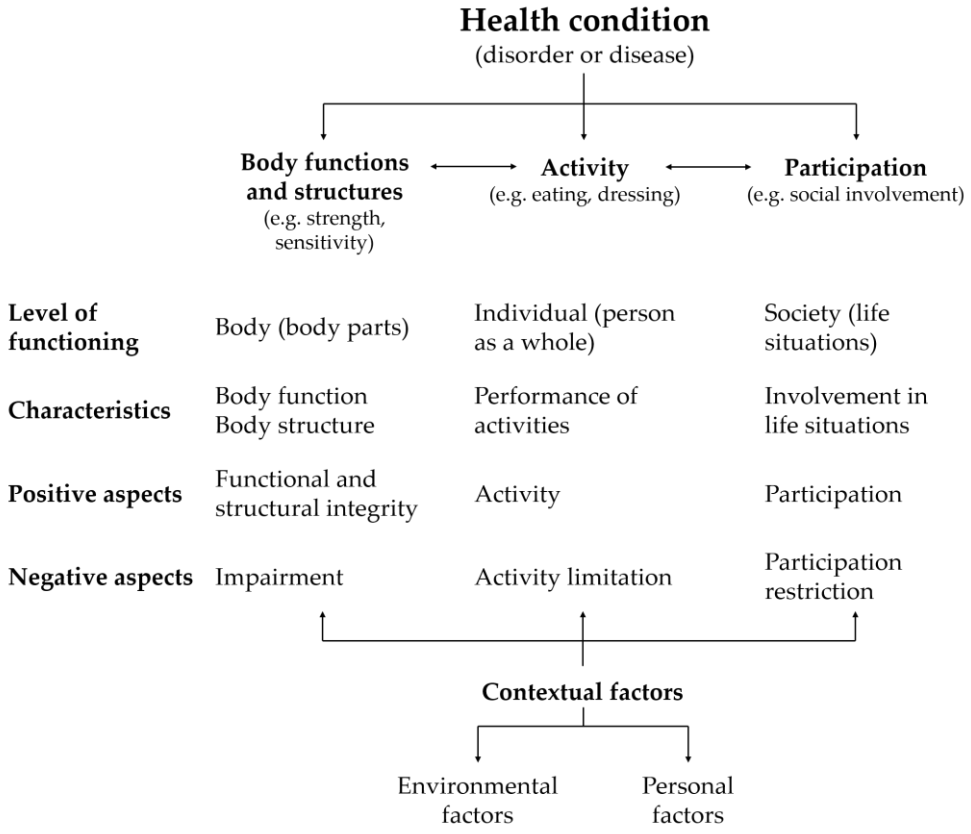

Figure 1.1. Overview of the dimensions of the ICF (adapted from World Health Organization, 2001).

Impairments are typically evaluated using clinical examination, or proven tools such as radiography to visualize bones, goniometry to measure range of motion or dynamometry to measure grip strength. However, these measurements do not inform the clinician about the performance of upper limb activities in everyday life (Barbier et al., 2003). Activity limitations are more complex to quantify. If we ought to define a gold standard for measuring activities, we would observe the patients while they are performing manual activities in their domestic environment. However, this is not practical for physicians given the burden and complexity of real-life assessments, so the patient's ability is assessed via a proxy, or PROMs in this case. For example, the ABILHAND questionnaire was designed to measure

manual ability, which is the ability of a person to use his/her hands and upper limbs to perform manual activities of daily living. The scale has been validated in populations with rheumatoid arthritis (Durez et al., 2007), chronic stroke (Penta et al., 2001), pediatric cerebral palsy (Arnould et al., 2004), systemic sclerosis (Vanthuyne et al., 2009) and neuromuscular diseases (Vandervelde et al., 2010). The validation for its use in hand surgery is described in the next section.

1.3.2. The classical test theory

The classical test theory (CTT) is a traditional quantitative approach to testing the reliability and validity of a questionnaire based on its items. Items are scored according to a rating scale (e.g. Impossible = 0, Difficult = 1, Easy = 2), and scores to each item are then summed up to generate a total questionnaire score. The CTT framework focuses on the questionnaire as a whole. It is based on the idea that a person's observed score on a test is the sum of a true score (error-free score) and an unsystematic (i.e. random) error (Spearman, 1904).

True scores quantify the latent trait to be measured (i.e. the attribute of interest). As values of the true score increase, responses to items representing the same concept should also increase, assuming that item responses are coded so that higher responses reflect more of the underlying latent trait. Random errors found in the observed scores are normally distributed, and therefore, the mean of such random fluctuations is taken to be zero. Random errors are assumed to be uncorrelated with the true score.

The previously cited questionnaires (MHQ, DASH, PRWE and CTQ) have been developed using the classical test theory. The premise is that individual item scores can be summed up (without weighting or standardization) to produce a total score (Lord and Novick, 1968). The CTT

is widely used for questionnaire development, but has several limitations (Smith et al., 2002):

- Scores are not necessarily objective as item and test indices depend on the examined sample. For example, a more able group will score higher than a less able one on the same test, and a person will seem more able if an easy test is administered, compared to a more difficult one.

- The CTT cannot predict an individual response to a given item. For example, it is not possible to predict how a person who answered "easy" to 50% of the items on a questionnaire would rate the difficulty of an item that was answered "easy" by 80% of the patients taking the same questionnaire (Smith et al., 2002).

- People's level of function cannot be measured independently of the difficulty of the test used.

- Total scores obtained by adding up the values of each response are ordinal and not necessarily linear, which means that the measurement unit is not constant throughout the measurement range. The same distance between scores (e.g. from 0 to 1 and from 1 to 2) may not reflect the same amount of increase in ability. This distortion of the score is especially noticeable at the extremes of the score range, compared to the center of the scale. For example, a 2-point difference at the center may represent a smaller true score difference than a 2-point difference at the extremities (DeVellis, 2006). As a consequence of ordinal scores, many of the statistical models to make mathematical comparisons among individuals or groups are invalid, as these assume an interval scale (Merbitz et al., 1989; Wright and Linacre, 1989).

- The standard error of measurement is only know at sample-level, but not at subject-level. Hence, only the average error of measurement

across the whole sample is known, and assumed to be the same for all subjects, even though it is known that scores at the extremities are generally less precise. As a consequence, the CTT prevents an individual approach to assess functional change for patient follow-up. Using the CTT, functional change can only be quantified based on group-level indices such as the variance between subjects at a given point in time or the average change between assessments.

- Scores are not easily interpretable, nor comparable between patients unless the data is complete, and missing responses are difficult to manage.
- The CTT cannot validate response patterns. If a patient affirms that the easiest items are "difficult" and that the most difficult items are "easy", does their score truly reflect their ability level?

1.3.3. The Rasch model

When assessing variables in medicine, we can measure either observable or latent variables. Observable variables (i.e. physical features) such as grip strength or range of motion can be directly measured with an instrument such as a dynamometer for grip strength and a goniometer for range of motion. Latent variables (e.g. manual ability or intelligence) are variables that are not directly observed, but can be accessible to measurement if they manifest themselves through external physical events (Tesio, 2003). For example, intelligence can become manifest through problem solving, and manual ability through the execution of activities requiring the use of the hands. By observing and counting these observable events, we can deduce the amount of the latent variable that is concealed within the subject (e.g. the more activities can be accomplished, the higher the level of manual ability). However, counting does not provide cues to valid quantitative measurements. For this reason, we need a model that relates counts of

observations to an abstract linear continuum from "less" to "more" (Tesio, 2003). Measurement requires an abstraction or construct which represents the attribute being measured. When measuring an observable variable, such as the length of an object, we refer to an abstract continuum on the measurement instrument being used. A line conceptualized from "less" to "more" represents a gradient of increasing levels of the variable. Measurement of any variable should comply with the fundamental principles of measurement, as will be developed in this section: linearity, unidimensionality, invariance, and objectivity (Tesio, 2003).

The Rasch model (Rasch, 1980) is a statistical approach to measuring latent variables such as human performance, attitudes, and perceptions (Tesio, 2003). It was developed by the Danish mathematician Georg Rasch in the 1960s, and has become increasingly popular in health and human sciences as awareness of the limitations of the CTT has grown (Conrad and Smith, 2004; Smith et al., 2002). The measurement principles and methodological concepts underlying the Rasch model have been detailed in the referred textbooks (Andrich, 1988, 1978; Rasch, 1980; Thurstone, 1959; Wright and Stone, 1979; Wright and Masters, 1982). It is based on the assumption that patients with a higher level of the measured latent trait (manual ability in this case), will have a higher probability to successfully pass an item, compared to patients with lower ability levels. The model states that the probability to pass an item depends only on subject ability and item difficulty (Rasch, 1980), according to the formula

$$P(X = 1|0,1) = \frac{e^{\beta-\delta}}{1 + e^{\beta-\delta}}$$

where P is the probability of passing an item, β is the subject ability and δ is the item difficulty. Figure 1.2 illustrates the dichotomous Rasch model where each item has only two possible outcomes: pass or fail (i.e. able or unable to achieve the activity).

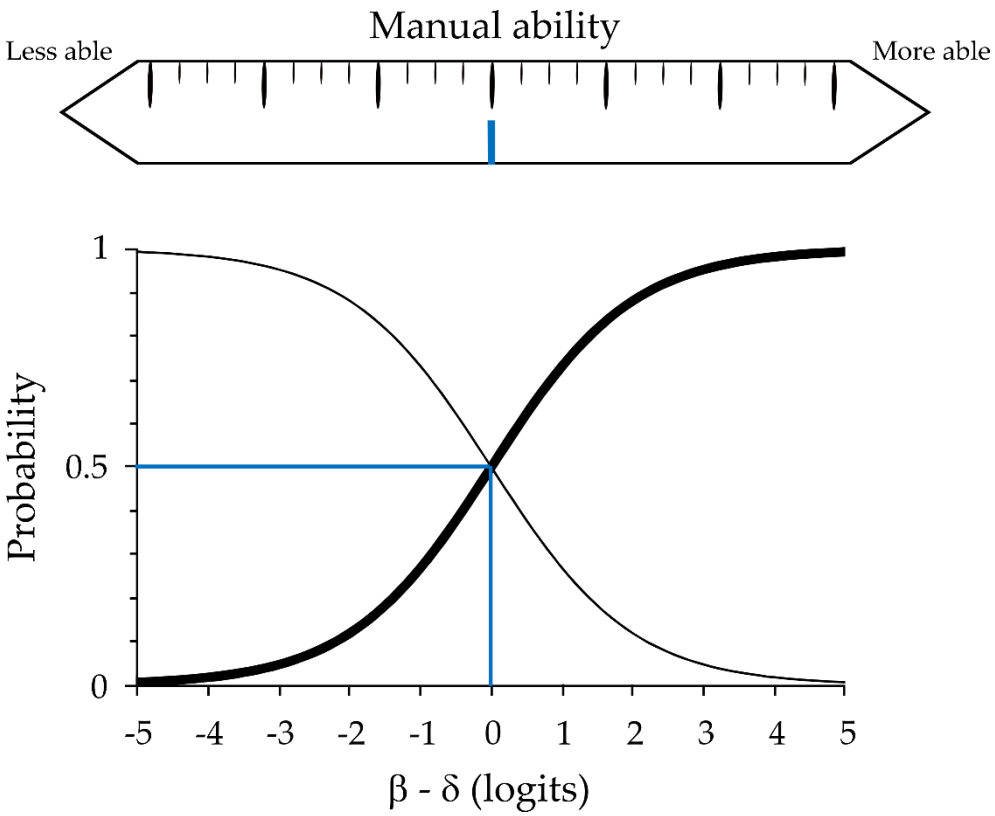

Figure 1.2. Probability of passing or failing an item in the dichotomous Rasch model. The thick and thin lines represent the probability of passing or failing an item, respectively, as a function of the difference between subject ability (β) and item difficulty (δ). The top panel illustrates the manual ability continuum. The blue tick on the ruler is positioned at 0 logits, for which β=δ, and the probability of passing or failing an item is 50%. The higher the subject ability, the higher the probability of success will be.

These two parameters can be estimated based on the proportions of responses to each item. For example, the item that got the highest proportion of "passes" over "fails" (the item that most subjects were able to pass) is the easiest one, while the item that got the lower pass/fail proportion is the most difficult one. Likewise, the subject that managed to pass most items is the most able one, while the one who failed most items is the least able one. Using the Rasch model, subjects and items can be placed on a common linear

scale. The latent variable "manual ability" can thus be conceptualized as an infinite continuum representing ability levels from "less able" to "most able". Measuring a patient's manual ability involves determining the patient location along this continuum.

The graduations of this scale are formed by the items of the questionnaire. As in a physical measuring instrument such as a ruler, the whole range of measurement should be covered by graduations in order to achieve the greatest precision.

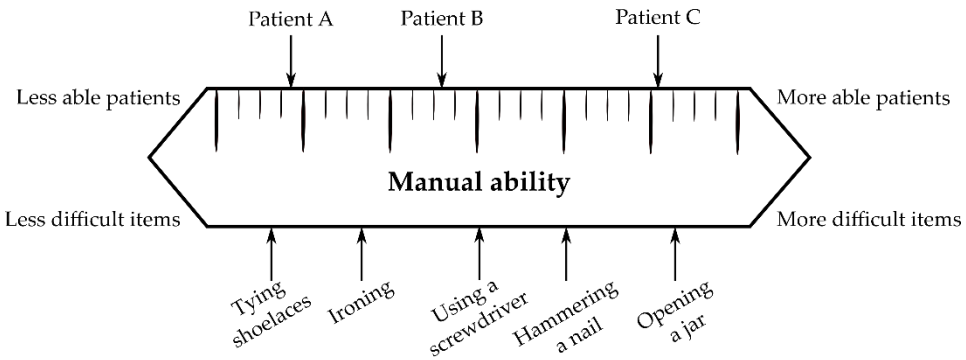

Figure 1.3. Manual ability continuum. Arrows represent patient (upper arrows) and item (lower arrows) locations on the continuum.

Figure 1.3 shows the continuum, represented by a ruler, along which the patients are located from the least to the most able, and the items from the easiest to the most difficult. As illustrated in Figure 1.3, patient A has a low manual ability level since his/her ability level is just enough to pass the first (easiest) item. Patient B has a moderate ability level and is expected to successfully pass the two easiest items and fail the three most difficult ones. Patient C has a high ability level and can likely succeed in all items except the most difficult one. As the Rasch model is probabilistic, exceptions can sometimes occur and a patient can potentially succeed in a difficult item while failing an easier one.

The Rasch model has been adapted to the polytomous response format, where an item can have different response categories such as impossible/difficult/easy (Andrich, 1978; Masters, 1982; Wright and Masters, 1982). These models state that the probability of response to an item depends only on the patient's ability, item difficulty, and threshold difficulties. Thresholds are located between two adjacent response categories, and correspond to the ability level needed for the patient to have a higher probability of selecting a particular response category rather than the previous adjacent one. In a polytomous response format (Figure 1.4) where activities can be answered on a three-level scale (impossible/difficult/easy), the thresholds between successive response categories are the graduations of the scale (compared to the items in the dichotomous model, since these only have one threshold). Therefore, patients whose manual ability is located beneath the first threshold are most likely to be unable to accomplish the activity; patients with an ability level located between the two thresholds are expected to perform the activity with difficulty, and patients with a manual ability located after the second threshold are most likely to complete the activity easily.

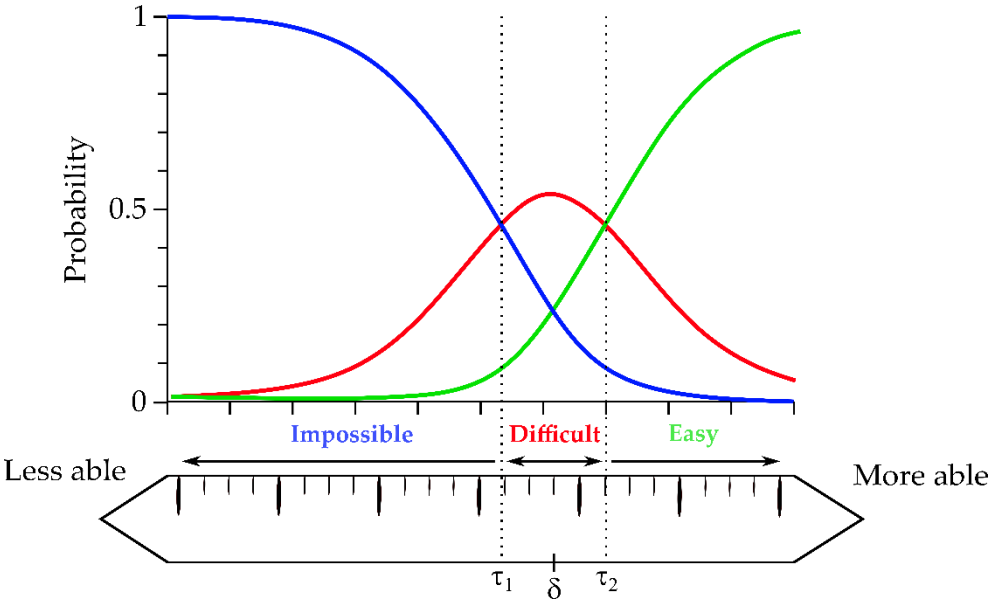

Figure 1.4. Polytomous response format. The blue, red and green lines represent the probability of answering "Impossible", "Difficult" or "Easy", respectively, to a given item as a function of manual ability. The larger the ability of a subject, the more probable he/she is likely to choose the higher response category. The intersection between two consecutive response categories corresponds to a threshold (τ), represented by the dotted lines. The first threshold (τ_1) corresponds to the ability level required to respond "difficult" rather than "impossible", while the second threshold (τ_2) is located at the ability level required to respond "easy" rather than "difficult". The item difficult δ corresponds to the mean of the two thresholds.

The more graduations are found on the scale, the more the measurement will be precise, as is the case with a ruler that is graduated each millimeter, compared to one tick each five millimeters (Tesio, 2003). When the measurement continuum is divided into more parts, the sensitivity to change and reliability are also expected to increase (Cano et al., 2006; Hobart et al., 2007). The more response categories constitute an item, the more the number of thresholds increases, and thus the more scale graduations per item. However, a certain balance must be achieved between increasing the scale precision and not confusing the respondent, which would increase the

measurement error. For example, as an answer to the question "How do you feel after surgery?", a patient can be presented with the following response categories: almost the same/a little better/somewhat better/moderately better/a good deal better/a great deal better/a very great deal better (Jaeschke et al., 1989). While the presence of seven categories in this Likert scale gives the illusion of increased precision, it might just confuse the patient who would not be able to discriminate between so many categories (Penta et al., 2001). The Rasch model can determine if the response categories are functioning as intended (i.e. they are well discriminated by patients). Successive response categories such as impossible/difficult/easy should represent increasing levels of ability. For a patient with a given manual ability level who performs an activity with difficulty, it is expected that a more able patient would answer "easy", and a less able patient would answer "impossible". The Rasch model investigates category functioning by verifying whether thresholds between adjacent categories are located at increasing levels of ability (i.e., that the thresholds are correctly ordered) (Andrich, 1996).

A very important property of a measuring instrument is unidimensionality, which means that the scale measures only one variable (or attribute of an object) without being influenced by other factors (Brentani and Golia, 2007). For example, a ruler measures only one property of an object (size) and is not influenced by other object properties such as shape and color. In the case of latent variables, the theoretical concept of unidimensionality is never totally met in practice, since the separation of one trait from the others is extremely difficult (Andrich, 1988). Approximating this ideal in the observed data is required so that subjects can be quantitatively compared based on the same attribute (manual ability in this case) (Wright and Linacre, 1989). Unidimendionality is tested by comparing the observed responses to an item with the expected responses predicted by the model. The differences between the two responses are compared using fit statistics reported by the different softwares used for Rasch analysis.

These statistics determine how closely the items define the underlying construct, and detect items that do not contribute to the definition of a uinidimensional scale of manual ability.

As by the definition of unidimensionality, the measured variable should not be influenced by other patients' characteristics such as gender or age. The invariance of the scale can be tested among subgroups of patients with differential item functioning (DIF) (Holland and Wainer, 1993). A DIF is present if for a given manual ability level, a subgroup of patients (e.g. male patients) find an item easier or more difficult than another subgroup (e.g. female patients). The presence of a DIF introduces a systematic misfit to a common scale calibrated for all subjects, and therefore restricts the use of the same scale for all subjects. The Rasch model can recognize DIF among subgroups and identify items presenting such bias. This allows the development of a common scale calibrated for all subjects.

A linear scale is obtained by converting ordinal raw scores into linear measures of the latent variable (manual ability). The units of the scale are "logits", a probabilistic unit that defines the pass/fail probability ratio for a patient to be able to achieve an activity: the higher the logit value, the higher the probability that a patient will manage an activity easily. This unit is constant along the entire range of the scale, which allows measures to be quantitatively compared and treated as a linear variable. Based on questionnaire raw scores, the model estimates a location for each patient (i.e. their ability) and for each item and threshold (i.e. their difficulty). All locations are scaled along a common linear, unidimensional continuum that defines the latent variable of interest (manual ability in this case). Each location has an associated standard error, which quantifies the degree of uncertainty associated with the estimated ability or difficulty. The standard error is not uniform across the range of the scale, but is generally smaller at the center and larger at the extremities of the scale. A good fit of the data with the model and the lack of DIF affirm invariant locations along the

continuum and indicate that the measures are unbiased with respect to patients' characteristics other than the one being assessed, i.e. manual ability.

Summary

The formulation of the Rasch model ensures that the resulting scale verifies the fundamental requirements of a measuring instrument: linearity, unidimensionality, invariance and objectivity (Merbitz et al., 1989; Rasch, 1980; Wright and Linacre, 1989). Linearity is ensured by the properties of the interval scale: the distance between scale graduations is constant throughout the range of measurement. Unidimensionality is verified by selecting the items that fit with the model; the model requires that only one patient attribute (i.e. their manual ability) determines the response probability. Invariance in the patient-item interaction is also confirmed by testing that the probability of observing a given response does not vary with patient factors (e.g. age, gender, level of education…) other than the one being measured (manual ability) (Rasch, 1980; Wright and Linacre, 1989). Objectivity is achieved when comparisons between individuals become independent of which particular instruments (e.g. questionnaires) have been used to generate the measures (Rasch, 1980). With these criteria met, the resulting questionnaire verifies the properties of a measuring instrument, as a ruler is used for size measurement. Just like one centimeter represents the same length throughout the range of any size measurement instrument **(linearity)**, the increase in subject manual ability by one logit corresponds to the same increase in manual ability by a constant factor of 2.71 (i.e. the Neperian constant) (Wright and Masters, 1982). The ruler measures only one property of the object (i.e. size) and is thus **unidimensional**. A multidimensional instrument that combines two properties of the object (e.g. size and weight) in a single score would be less intuitive to interpret. The measures of two objects obtained with the same ruler do not depend on other objects properties like shape and color (size measurement instruments are **invariant** relative to objects qualities other than their size). The size of the object remains the same whether the measure has been made with a ruler or

a measuring tape, hence the **objectivity** of the measure, or the independence of the measure relative to the instrument.

1.4. Thesis content

In this thesis, we explore different patient-centered tools that aim to improve outcome assessment in hand surgery. In this introduction (Chapter 1), we presented the general context of outcome measurement and current tools and theories to evaluate the domains of interest. In Chapter 2, we describe the validation of the ABILHAND questionnaire for hand surgery. In Chapter 3, we illustrate how the developed questionnaire can be used in clinical practice. In Chapter 4, we study the responsiveness and minimal clinically important difference of the ABILHAND-HS. In Chapter 5, we present the prototype of a device that could be used to monitor manual activities of daily living. We also describe an algorithm that classifies manual activities into different categories. In Chapter 6, we summarize the contributions of this thesis, we identify the future directions that could follow the preliminary results of this work, and we discuss the potential contributions of the work to the field of outcome measurement.

CHAPTER 2

Manual ability in hand surgery patients: validation of the ABILHAND scale in four diagnostic groups

Published as: El Khoury G, Barbier O, Libouton X, Thonnard JL, Lefèvre P, Penta M. Manual ability in hand surgery patients: Validation of the ABILHAND scale in four diagnostic groups. PLoS One. 2020 Dec 3;15(12):e0242625.

Patients treated in hand surgery (HS) belong to different demographic groups and have varying impairments related to different pathologies. HS outcomes are measured to assess treatment results, complication risks and intervention reliability. A one-dimensional and linear measure would allow for unbiased comparisons of manual ability between patients and different treatment effects. A preliminary 90-item questionnaire was presented to 216 patients representing the diagnoses most frequently encountered in HS, including distal radius fracture (n=74), basal thumb arthritis (n=66), carpal tunnel syndrome (n=53), and heavy wrist surgery (n=23). Patients were assessed during the early recovery and in the late follow-up period (0-3 months, 3-6 months and >6 months), leading to a total of 305 assessments. They rated their perceived difficulty with queried activities as impossible, difficult, or easy. Responses were analyzed using the RUMM2030 software. Items were refined based on item-patient targeting, fit statistics, differential item functioning, local independence and item redundancy. Patients also completed the QuickDASH, 12-item Short Form Survey (SF-12) and a numerical pain scale. The rating scale Rasch model was used to select 23 mostly bimanual items on a 3-level scale, which constitute a unidimensional, linear measure of manual ability with good reliability across all included diagnostic groups (Person-Separation Index = 0.90). The resulting scale was found to be invariant across demographic and clinical subgroups and over time. ABILHAND-HS patient measures correlated significantly ($p<0.001$) with the QuickDASH ($r=-0.77$), SF-12 Physical Component Summary ($r=0.56$), SF-12 Mental Component Summary ($r=0.31$), and pain scale ($r=-0.49$). ABILHAND-HS is a robust person-centered measure of manual ability in HS patients.

2.1. Introduction

In hand surgery (HS), as in other medical specialties, outcome evaluations are needed to assess the effectiveness and reliability of the intervention, as well as to reinforce patient education regarding the risks and outcomes of the procedure and, potentially, to justify therapeutic practices to payers (Dubert, 2014). Physician-documented reports of HS outcomes based on clinical examination and imaging should be complemented with patient reported outcomes assessed by questionnaires designed to capture patients' perspectives with respect to the impact of their conditions and interventions on their daily lives (Berkanovic et al., 1995; Hewlett, 2003; Ziebland et al., 1993).

Current views of health and disability have been shaped by the World Health Organization's International Classification of Functioning, Disability, and Health (World Health Organization, 2001), which parses disease consequences into three domains: impairment of anatomical structures (e.g. bones, muscles, ligaments) or body functions (e.g. motor skills, sensitivity), activity limitations (e.g. manual activities), and participation restrictions (e.g. in hobbies and work). The impact of a pathology or a surgical intervention in these three domains is also conditioned by personal factors (motivation, capacity to develop compensatory strategies) and environmental factors (social or professional context). Although impairment measurements such as imaging can provide clues regarding functional prognosis, it does not provide good information about performance in everyday life, especially of the hands, which are important for a great variety of activities (Barbier et al., 2003; Bobos et al., 2018; Penta et al., 2001). For example, demonstration of a bone fracture union is insufficient to determine whether a patient is capable of resuming their usual activities or returning to work (Giladi and Chung, 2013; Jaremko et al., 2007; Synn et al., 2009; Young and Rayan, 2000).

The patient-reported questionnaires that have been most commonly used in HS (Changulani et al., 2008) are the Disability of the Arm, Shoulder

and Hand questionnaire (DASH) (Hudak et al., 1996), the Patient Rated Wrist Evaluation (PRWE) (MacDermid et al., 1998), and the Carpal Tunnel Questionnaire (CTQ) (Levine et al., 1993). Each of these questionnaires has been reported to have good psychometric properties, but each has a particular focus on its own area(s) of disablement. The DASH assesses body functions, activities, and participation (Coenen et al., 2013) and can be divided into 3 subscales based on dimensionality (Franchignoni et al., 2010). Meanwhile, the PRWE is specific to the wrist joint and the CTQ is specific to carpal tunnel syndrome (CTS). The Michigan Hand Outcomes Questionnaire (MHQ) (Chung et al., 1998) is a multidimensional hand-specific outcomes instrument consisting of six subscales, measuring overall hand function, activities of daily living, pain, work performance, aesthetics, and satisfaction. It measures impairment by hand (left and right separately), rather than overall disability. Interpretation of total scores on multi-dimensional instruments can be less than straightforward given that patients can show simultaneous improvement in one domain with deterioration in another (Merbitz et al., 1989; Wright and Linacre, 1989). Assessment of functional recovery on a unidimensional (Thurstone, 1959) and linear (Wright and Linacre, 1989) scale would allow for quantitative comparisons of ability among different patients and treatments. Such a scale can be developed with state-of-the-art psychometric methods, such as the Rasch model (Grimby et al., 2012; Rasch, 1980).

The ABILHAND questionnaire is a Rasch-model built measure of manual ability (Penta et al., 1998) that provides an invariant linear scale and allows for quantitative comparisons of manual ability between patients and over time. The scale has been validated in populations with rheumatoid arthritis (Durez et al., 2007), chronic stroke (Penta et al., 2001), pediatric cerebral palsy (Arnould et al., 2004), systemic sclerosis (Vanthuyne et al., 2009) and neuromuscular diseases (Vandervelde et al., 2010). These previous validations have shown that the difficulty of most manual activities was diagnosis-dependent (Arnould et al., 2012). Therefore, the objective of this

work was to adapt the ABILHAND scale to the most frequent diagnoses treated in HS.

2.2. Methods

2.2.1. Questionnaire adaptation to HS patients

The ABILHAND is a measure of manual ability that assesses one's ability to manage daily activities requiring upper limb use, regardless of strategy (Penta et al., 1998). The necessary permissions were obtained from the developer of the original questionnaire to modify it. To develop a HS-adapted ABILHAND, a preliminary item list was compiled from previous versions of the ABILHAND questionnaire, the DASH, PRWE, CTQ, and MHQ items together with some new items. This pool of items was submitted to nine HS experts (hand surgeons, physical medicine and rehabilitation physicians, physical therapists, and occupational therapists), who were asked to assess each item's relevance to hand surgery patients on a yes/no basis and propose additional items that might be affected by the relevant pathologies (e.g. sensation for CTS and wrist loading for distal radius fractures (DRF)). A final list of 90 items constituted the experimental ABILHAND-HS questionnaire.

2.2.2. Patients

A convenience sample of 216 patients was recruited from February 2018 to February 2019 at the HS consultation center at Cliniques Universitaires Saint-Luc, Belgium representing the following four diagnostic categories: CTS, DRF, basal thumb arthritis (BTA), and heavy wrist surgery (HWS, including 1st row carpectomy and partial or total wrist arthrodesis). The inclusion criteria for patients were being >18 years old and being able to read

and understand French. The exclusion criteria included comorbidities that may impede manual ability substantially (i.e. tremor, paralysis and active rheumatologic disease) and any mental or cognitive dysfunction (i.e. dementia and mental retardation). The patient characteristics are summarized in Table 2.1. Patients provided written informed consent to participate. This study was approved by the ethical committee of Cliniques Universitaires Saint-Luc-Université catholique de Louvain (N° B403201523492).

Table 2.1. Sample characteristics (n = 216).

Characteristic	N (%)[a]
Gender	
Women	145 (67%)
Men	71 (33%)
Mean age (range), years	60.3 (19–93)
Education level	
Basic	109 (51%)
Postsecondary	107 (49%)
Work status	
Student	2 (1%)
Unemployed	22 (10%)
(Self-)Employed	83 (38%)
Retired	109 (51%)
Hand dominance	
Right	194 (90%)
Left	15 (7%)
Ambidextrous	7 (3%)
Involved dominant hand	
Yes	136 (63%)
No	80 (37%)
Diagnostic group	
Distal radius fracture (DRF)	74 (34%)
Basal thumb arthritis (BTA)	66 (31%)
Carpal tunnel syndrome (CTS)	53 (24%)
Heavy wrist surgery (HWS)	23 (11%)
Follow-up assessments (n = 305)	
0–3 months (57 DRF, 52 CTS, 22 BTA, 1 HWS)	132 (43%)
3–6 months (38 DRF, 16 CTS, 3 BTA, 1 HWS)	58 (19%)
>6 months (30 DRF, 18 CTS, 46 BTA, 21 HWS)	115 (38%)

[a]**Except where otherwise indicated**

2.2.3. Procedures

The French-language experimental ABILHAND-HS items were presented in five random orders to avoid a systematic item sequence bias. Patients were asked to indicate their perceived difficulty associated with completing the activities without technical or human assistance, independent of the hand used to perform the activity on a three-level scale: impossible (0), difficult (1), or easy (2) (Penta et al., 2001). Activities not attempted during the last week were treated as missing responses. Patients also completed the QuickDASH (Beaton et al., 2005), 12-item Short Form Survey (SF-12) questionnaire (Ware et al., 1996) and a 10-level numerical pain scale, for external validation purposes.

Patients were first assessed as soon as they presented to their hand surgery consultation appointments and had experienced manual activities in their own environment: after hand surgery and cast removal for DRF, BTA and HWS and at the first consultation for non-operated CTS and BTA. For the first assessment, patients were interviewed by the principal investigator in order to ensure clarity, obtain feedback from participants, and make sure instructions are properly followed. Patients were also asked to suggest additional items they felt the questionnaire was missing. However, these were either gender related (e.g. fastening a bra) or very specific and were thus not retained. Follow-up assessments were completed in our consulting office or returned by mail, leading to a total of 305 completed assessments, which provides sufficient power to support the planned Rasch analysis (Hagell and Westergren, 2016).

2.2.4. Rasch analysis

The 90-item experimental ABILHAND-HS questionnaire responses were analyzed using the Rasch model in RUMM2030 software (RUMM Laboratory Pty Ltd., Perth, Australia). The Rasch model (Rasch, 1980), a

prescriptive model, requires that specified response probabilities depend on only item difficulty and patient ability. Polytomous datasets with thresholds between successive response categories can be analyzed with either a rating scale model that constrains all threshold locations to be equal across items (Andrich, 1978) or a partial credit model that allows threshold locations to vary across items (Masters, 1982). Patient abilities and item difficulties are located along a common linear, unidimensional continuum that defines the latent variable of interest (i.e. manual ability). The locations are expressed in logits, calculated as the logarithm of the pass/fail probability ratio of an item or threshold. The logit locations were converted into centiles to facilitate clinical interpretation on a linear scale ranging from 0% (smallest ability) to 100% (largest ability) (van Nes et al., 2011). Expected responses, determined based on the patient and item locations, were compared to the responses actually reported to compute residual and fit statistics, which were then used to assess the scale's unidimensionality (Andrich et al., 2013). A good fit of the data with the model affirms invariant locations along the continuum and indicates that the measure can be used to compare manual ability across patients and diagnoses.

2.2.5. Item selection

From the experimental version of the questionnaire, the ABILHAND-HS was refined through successive analyses of 305 assessments with the goal of selecting items that define a unidimensional and clinically relevant scale of manual ability. P values < 0.05 were considered significant for each of the following analysis steps:

1) *Item-patient targeting.* Based on examination of patient distributions and item locations, items that showed a floor effect (too easy) or did not target the patients sample ability were removed.

2) *Rating scale*. Items with disordered thresholds and items with thresholds that were too narrow (<1.4 logits) or too wide (>5 logits) were removed before applying the rating scale model (Linacre, 2002).

3) *Unidimensionality*. Only items that delineated a common manual ability construct according to the following four criteria were retained: (1) standardized residuals obtained over three class intervals had to be within ±2.5 with a non-significant χ^2 (Andrich et al., 2013); (2) no observable major differential item functioning (DIF) (Holland and Wainer, 1993), uniform or non-uniform, shown by a 2-way analysis of variance of the residuals with Bonferroni correction (Armstrong, 2014), according to gender (male vs. female), age (above vs. below the median age of 63 years), pathology (CTS vs. DRF vs. BTA vs. HWS), involved hand (dominant vs. non-dominant), level of education (basic vs. superior), and follow-up (0-3 months vs. 3–6 months vs. >6 months); (3) overall fit of the response set based on a non-significant item-trait interaction χ^2 (Andrich et al., 2013); and (4) statistically similar patient locations, according to paired t-tests, calculated with items that loaded either positively or negatively on the first residuals principal component (Linacre, 1998; Pallant and Tennant, 2007; Smith, 2002).

4) *Local independence.* When items were found to be querying redundant content (Wright B.D., 1996), demonstrated by a residual correlation > 0.3, the item with the poorer fit statistic was deleted (Ramp et al., 2009).

5) *Item redundancy*. To shorten the scale, when two or more items had similar locations on the continuum, the one with the best fit was retained.

2.2.6. Scale reliability

The Person-Separation Index (PSI), i.e. the proportion of total variance (including error) that is attributed to patient location variance, was used to determine the ABILHAND-HS scale's reliability and its degree of precision with the dataset, and thus how many statistically different ability strata can be distinguished along the scale (Fisher, 1992).

2.2.7. Construct validity

The construct validity of the ABILHAND-HS was examined with a comparison of means for associations with gender, involved hand, and diagnosis. The relationships of the ABILHAND-HS with age, the QuickDASH scale, the numerical pain scale, the SF-12 Physical Component Summary (PCS), and the SF-12 Mental Component Summary (MCS) were assessed with a correlation analysis.

Patient perceptions were compared between ABILHAND-HS and QuickDASH items by adding the six QuickDASH activity items to the anchored data matrix. The locations of similar items were then compared between the scales.

2.2.8. Statistical analyses

Statistical analyses were completed in IBM SPSS Statistics for Windows, version 25 (IBM Corp., Armonk, N.Y., USA). Data normality was verified for statistical tests using the Shapiro-Wilk test and Q-Q plots. Parametric tests were used for normal data and continuous variables, non-parametric tests for non-normal data and ordinal variables. A Mann-Whitney u-test (two-tailed) was used for gender differences, an independent-samples t-test (two-tailed) for association with the involved hand, and an analysis of variance for diagnosis. Pearson correlation coefficient was calculated for association

with age, while relationships with the QuickDASH scale, the numerical pain scale, the SF-12 Physical Component Summary (PCS), and the SF-12 Mental Component Summary (MCS) were assessed with Spearman correlation coefficients. P values < 0.05 were considered significant. Mean values are reported with standard deviations (SD). Chi-square and t values are reported with degrees of freedom (df).

2.3. Results

2.3.1. Item selection for the ABILHAND-HS scale

Successive analyses led to the selection of 23 items defining a unidimensional manual ability scale in HS. Of the 90 experimental items, 34 were removed because they were too easy (e.g. 'Drinking a glass of water'), 3 items had too-narrow thresholds (e.g. 'Using a touch screen'), 4 items were misfitting (e.g. 'Carrying a shopping bag'), and 26 items had a location redundant with another better fitting item (e.g. 'Peeling onions' was deleted in favor of 'Peeling potatoes with a knife').

2.3.2. Metric properties

The calibration obtained for the 23 mostly bimanual activities retained for ABILHAND-HS is reported in Table 2.2 in descending difficulty order. The standardized residuals obtained matched the expected standard normal distribution for items [mean (SD), -0.30 (0.99)] and for patients [0.31 (0.97)], indicating that the ABILHAND-HS scale is globally unidimensional. An invariant item location was obtained for more- and less- able patients as shown by a nonsignificant item-trait interaction ($\chi^2 = 57.76$, 46 df, $p = 0.11$). An invariant patient ability was obtained with items with different content as shown by a non significant t-test when using items that loaded positively

or negatively on the first principal residual component (t = 1.24, 304 df, p = 0.22).

Table 2.2. Calibration of the 23 items of the ABILHAND-HS.

Item	Bi-manual	Difficulty logits (centiles)	SE logits	Residual z	Fit χ^2	*P*
a. Doing push-ups	x	3.54 (78)	0.21	0.61	0.46	0.79
b. Playing a racket sport	x	2.30 (68)	0.25	0.04	1.66	0.44
c. Cutting a hedge	x	2.00 (65)	0.18	0.05	3.83	0.15
d. Opening a screw-topped jar	x	1.30 (59)	0.12	0.48	6.19	0.05
e. Applauding vigorously	x	1.11 (58)	0.16	2.09	2.00	0.37
f. Lifting a full pan	x	0.96 (57)	0.12	-0.53	5.69	0.06
g. Wringing a towel	x	0.86 (56)	0.12	-2.00	0.14	0.93
h. Opening a can with a can opener	x	0.76 (55)	0.12	-1.83	0.49	0.78
i. Hammering a nail	x	0.45 (52)	0.16	-0.70	5.31	0.07
j. Shaking bed sheets	x	0.13 (50)	0.17	-1.95	3.49	0.17
k. Using a screwdriver	x	-0.05 (48)	0.14	-0.14	0.54	0.77
l. Peeling potatoes with a knife	x	-0.16 (47)	0.13	-0.75	5.75	0.06
m. Ironing	x	-0.38 (45)	0.15	-0.44	3.12	0.21
n. Taking the cap off a bottle	x	-0.52 (44)	0.13	-0.32	2.75	0.25
o. Cutting one's nails	x	-0.55 (44)	0.13	-0.05	0.33	0.85
p. Shuffling and dealing cards	x	-0.77 (42)	0.16	-1.21	6.18	0.05
q. Wiping windows		-0.77 (42)	0.15	0.01	1.29	0.52
r. Tying shoelaces	x	-0.96 (41)	0.14	-0.61	2.02	0.36
s. Tearing open a pack of chips	x	-1.16 (39)	0.15	1.70	1.31	0.52
t. Fastening the zipper of a jacket	x	-1.45 (37)	0.14	0.32	1.58	0.45
u. Turning a car steering wheel	x	-1.69 (35)	0.18	-0.26	0.78	0.68
v. Putting on gloves	x	-2.24 (30)	0.19	-0.80	1.99	0.37
w. Spreading butter on a slice of bread	x	-2.68 (26)	0.18	-0.60	0.86	0.65

Analysis of DIF of the ABILHAND-HS with six criteria yielded only four instances of uniform DIF among the 23 items (Table 2.3). A small magnitude DIF was revealed among diagnoses (Figure 2.1) with no substantial impact on scale invariance, as evidenced by a good overall fit. Note that items were not specifically calibrated to the HWS group because of a limited sample size (n=23) (Chen et al., 2014). No DIF was observed between the first and last assessments, showing satisfactory invariance to support the scale follow-up stability. Likewise, an intraclass correlation coefficient across the first and last assessments was 0.94, indicating excellent item-difficulty-hierarchy consistency and providing confidence for data pooling over different time points (Chang and Chan, 1995). The PSI in this sample was equal to 0.90, indicating the distinguishability of four strata of manual ability (Fisher, 1992).

Table 2.3. Differential item functioning (DIF) summary.

Label	Person factor	Magnitude (logits)	Type	Difficulty
Taking the cap off a bottle	Gender	1.36	Uniform	Women > Men
Opening a screw-topped jar	Gender	0.93	Uniform	Women > Men
Opening a screw-topped jar	Diagnosis	0.61	Uniform	BTA > CTS > DRF
Using a screwdriver	Involved hand	1.26	Uniform	Non-dominant > Dominant hand involved

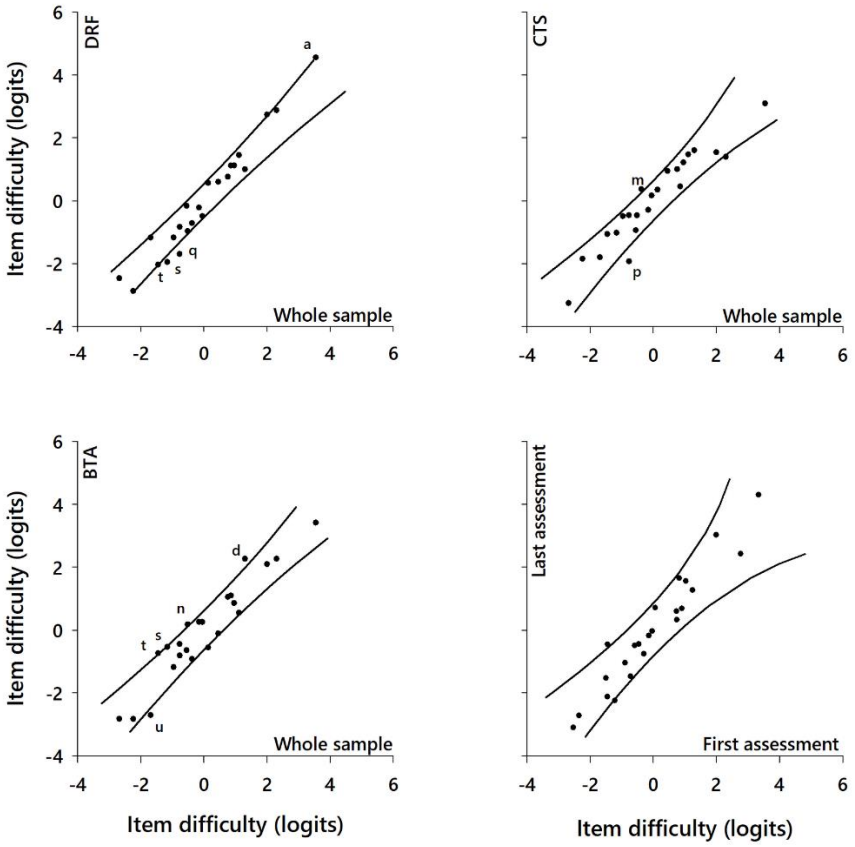

Figure 2.1. Differential item functioning (DIF) plots comparing the item difficulty hierarchy between subgroups. In each plot, the lines represent the 95% confidence interval of an ideal invariance between subgroups; the items are represented by the dots or by their letter if they display significant DIF. The most difficult items (dots) are plotted in the top right part of each plot. When comparing the item difficulty hierarchy between each diagnostic group relative to the whole sample, most of the ABILHAND-HS items lie within 95% confidence interval of the ideal invariance, indicating an invariant difficulty across diagnostic groups. When comparing the item difficulty hierarchy between the first and last assessment, all items fall within the 95% confidence interval of an ideal invariance, affirming invariance of item difficulties between the assessments.

2.3.3. Scale description

The ABILHAND-HS structure and targeting of HS patients are illustrated in Figure 2.2, showing an average patients' manual ability of 1.17 logits (SD = 1.85 logits; i.e. 58 (15) centiles). Twenty-four patients (7.9%) were able to perform all 23 activities easily, and were thus identified as extreme patients. Extreme patients tended to be younger men evaluated more than 6 months after treatment, and were more likely to have a CTS rather than a HWS. The three response categories were well distinguished in HS patients, with an inter-threshold distance of 2.93 logits (24 centiles), indicating that, regardless of patient ability, rating an item as 'easy' is about 20 (i.e. $e^{2.93}$=18.7) times more difficult than rating it as 'impossible'. Although the threshold distribution (range, -4.15 to 5 logits) was well targeted to the range of patient abilities, the patients' ability levels skewed high, indicating that the scale could measure patients that are more severely disabled than in this sample.

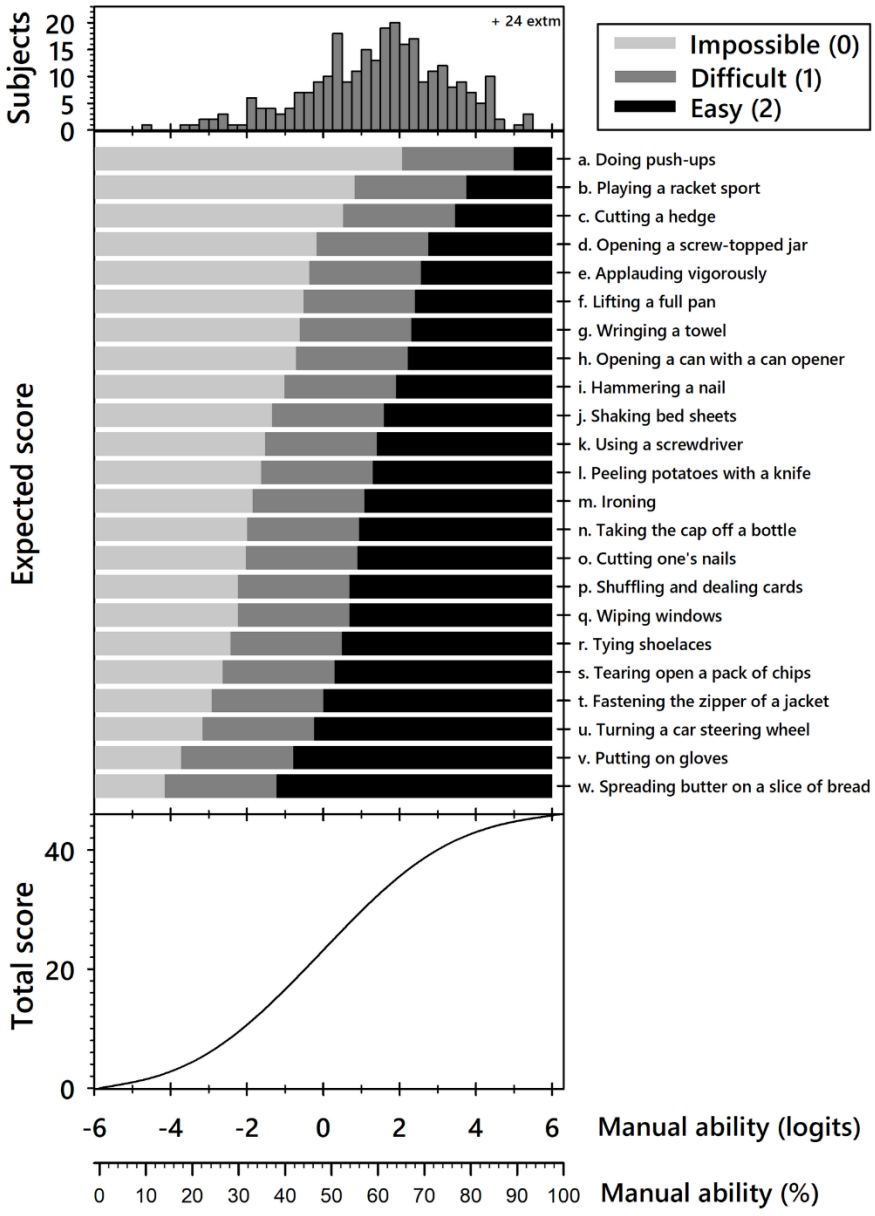

Figure 2.2. Structure of the ABILHAND-HS scale. Top: distribution of manual ability measures for the whole sample expressed in logits (log of the pass/fail probability ratio) and centiles (fraction of the measurement range). Twenty-four

patients (7.9%) were able to perform all 23 activities easily, and were thus identified as extreme patients. None of the participants reported that they could not perform any of the 23 activities. Middle: most probable patient response to each item based on the patient manual ability and on the difficulty of the item's response category. The average item difficulty was set to 0 logits and the items are ordered from most (top) to least (bottom) difficult. The distance between thresholds (middle bar) is constant for all items (2.93 logits or 24 centiles). A patient with a manual ability measure of 0 logits would be expected to perform the first 3 activities easily, to have some difficulty with the following 17 activities, and to be unable to perform the 3 most difficult activities. A patient with a measure of 2.1 logits should be able to perform all activities easily or with some difficulty. Bottom: conversion of ordinal raw scores into a linear continuum of manual ability for complete response sets. The raw scores ranged from 0 to 46 (sum of scores of 0–2 for 23 items). This curve is linear in its central (30^{th}~70^{th} percentile) range, with sigmoid flattening outside the central range, highlighting a non-linear relationship, especially at the extremities of the score range.

2.3.4. Construct validity

ABILHAND-HS measures were normally distributed across the whole sample and subgroups, except for men (W = 0.97; 100 df; p = 0.038). An effect of gender on ABILHAND-HS manual ability measures was observed, with men [1.88 (2.39) logits; median 1.74 logits] reporting a significantly higher mean manual ability than women [1.32 (1.93) logits; median 1.4 logits; U = 8642; p = 0.026]. Manual ability was not found to be significantly associated with age (R = -0.04; p = 0.47), the hand involved (t = 0.96; 303 df; p = 0.37), or the patient's diagnosis (F = 1.92; 3 df; p = 0.12). Although variance across diagnosis groups was not significant, we did observe a broad spectrum of manual ability. Patients with CTS reported the highest manual ability [1.9 (2.0) logits], followed by patients with BTA [1.5 (2.3) logits], DRF [1.4 (2.1) logits], and HWS [0.9 (1.8) logits].

The relationships between ABILHAND-HS measures and scores obtained with other instruments are shown in Figure 2.3. Briefly, ABILHAND-HS correlated strongly with QuickDASH scores, moderately with SF-12 PCS scores and pain scale scores, and weakly with SF-12 MCS scores. We observed substantial similarity with respect to manual ability scale locations between the ABLHAND-HS and QuickDASH activity items (Figure 2.4).

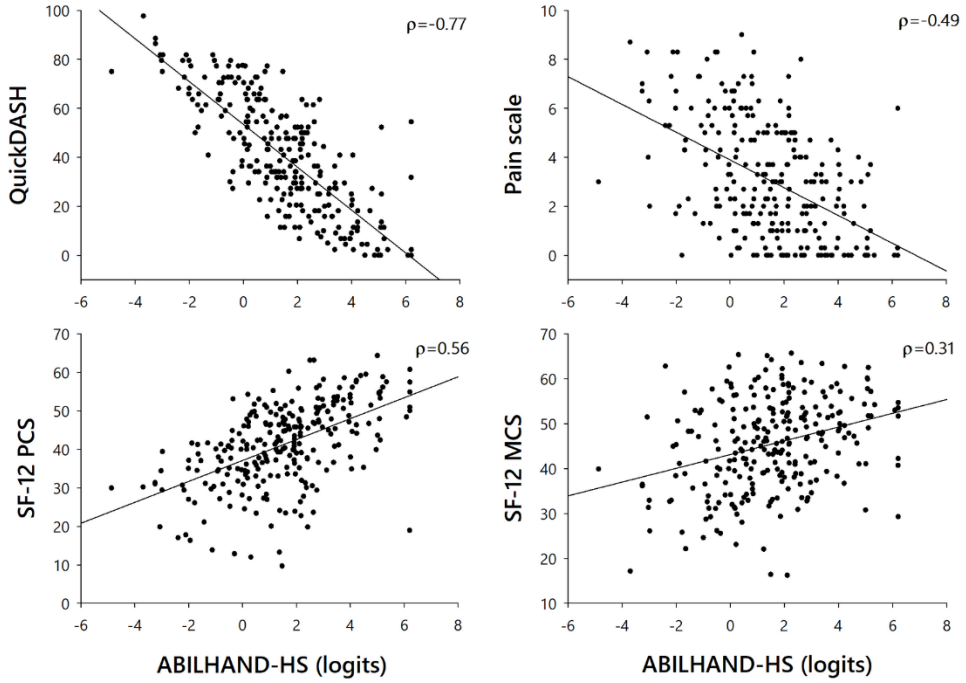

Figure 2.3. Correlations of ABILHAND-HS scores with QuickDASH, PCS, MCS, and numerical pain scale scores. Spearman correlation coefficients are indicated in the top right of each graph. All correlations were statistically significant ($p < 0.001$).

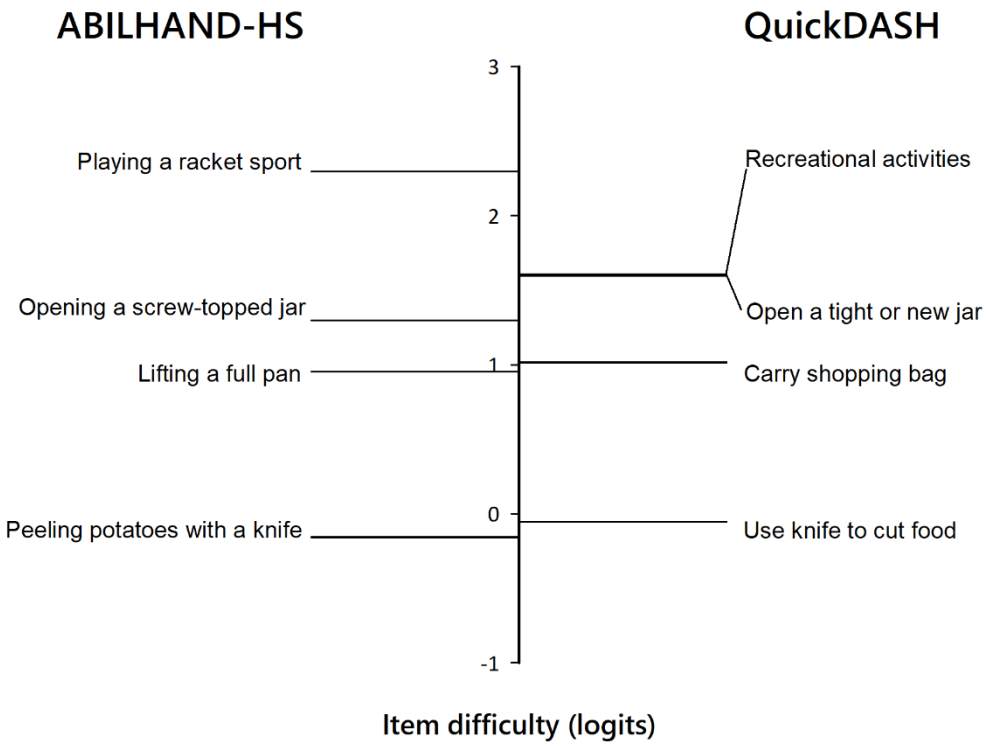

Figure 2.4. Comparison of difficulty levels (vertical axis) between similar items of the ABILHAND-HS (left) and QuickDASH (right) scales. QuickDASH item responses were added to the anchored data matrix of ABILHAND-HS responses to equate both measures.

2.4. Discussion

Here, we report the adaptation and validation of an ABILHAND-HS questionnaire for use with HS patients. Impairments present in our study cohort included weakness (e.g. following DRF), loss of sensation (e.g. in CTS), and stiffness (e.g. in BTA), with some patients presenting with a combination of these impairments. The ABILHAND-HS was constructed to measure manual ability on a common, linear, and unidimensional scale wherein the 23 activities retained delineate an invariant item difficulty

hierarchy independent of patient diagnosis. All ABILHAND-HS activities with the exception of one involve both hands and, consistent with our clinical experience, the most difficult ones require high levels of force (e.g. 'doing push-ups' loads the wrist in extension). Of the experimental 90 items, those that could be interpreted in different ways, for instance using the injured or uninjured hand, were misfitting and thus omitted (e.g. 'carrying a shopping bag'). The sample size was adequate for the statistical interpretation of fit statistics (Hagell and Westergren, 2016), and was within the same range of studies dealing with the development of outcome measures (Beaton et al., 2005; Chung et al., 1998; Hudak et al., 1996; MacDermid et al., 1998). The fit statistics for the 23 retained items support the item hierarchy invariance across the latent trait (Tennant and Conaghan, 2007). A few instances of minor DIF were retained to maintain the scale's construct validity (Hagquist and Andrich, 2017). The resulting scale is well targeted to the studied HS population, despite a small persistent ceiling effect, most likely due to missing responses for the most difficult activities. This observation of apparent ceiling effect involves 7.9% (24/305) of the records, which is well below the maximum recommended allowance of 15% (McHorney and Tarlov, 1995).

Although reliability indices should be compared with caution across potentially different study conditions, it is noteworthy that the PSI obtained for the ABILHAND-HS (0.90) was higher than prior values obtained for the activities subscales of the PRWE (Esakki et al., 2018) (0.78 and 0.81 for the usual and specific activities subscales, respectively, in DRF patients), for the Patient-Rated Wrist and Hand Evaluation (Packham and MacDermid, 2013) (0.83 in HS patients), for the QuickDASH scale (Franchignoni et al., 2011) (0.84 in patients with various upper limb dysfunctions) and for the Manual Ability Measure (Chen et al., 2005) (MAM-16; 0.83 for HS patients), while being equal to that for the DASH manual functioning subscale (Franchignoni et al., 2010). PSI values reflect sensitivity to clinical evolution over time, with greater values indicating a greater number of distinguishable ability strata.

We obtained person separation among patients using three response levels (impossible, difficult, and easy), consistent with previous studies showing patients unable to discriminate more than three levels of difficulty (ABILHAND (Penta et al., 2001), DASH activity items (Franchignoni et al., 2010) and QuickDASH activity items (Franchignoni et al., 2011)).

Accurate communication of scale administration instructions is critical for targeting patient manual ability as defined by the ABILHAND-HS. Generally, patients focus on their ability to perform the queried activities with their injured hand; likewise, the PRWE explores use of the affected hand explicitly (MacDermid et al., 1998). The ABILHAND-HS, like the QuickDASH, is oriented towards real daily life behaviors and is intended to be independent of the limb(s) or strategy used and unbiased by activities that are never performed with the affected hand or avoided during recovery (Penta et al., 2001). Our findings of stable item calibrations and lack of DIF across the assessments indicate that the ABILHAND-HS can be used confidently to assess the patient recovery at different time points during follow-up. Moreover, the stability of items hierarchy between the first and last evaluation indicate that the results were not influenced by the method of administration (interview with the investigator versus self-reported).

ABILHAND-HS construct validation results fit well with our clinical observations. The patients with the highest manual ability scores on the ABILHAND-HS also had the highest SF-12 PCS and SF-12 MCS scores as well as the lowest QuickDASH and numerical pain scale scores, which was also observed in other validation studies (Chung et al., 1998; MacDermid et al., 1998). Correlations of ABILHAND-HS with other instruments, including the QuickDASH, SF-12 and a pain scale are also consistent with prior findings suggesting that generic instruments are less sensitive than specific ones (Aktekin et al., 2011). The present ABILHAND-HS manual ability scores were not related significantly to age, consistent with other versions of the ABILHAND (Durez et al., 2007; Penta et al., 2001; Vandervelde et al., 2010). Our findings of a small, but significant gender effect, with men

tending to report a higher manual ability than women (mean difference, 0.56 logits), varied across HS diagnoses but, generally, were consistent with previous reports in patients with DRFs and wrist arthrodesis (Amorosa et al., 2011; Cowie et al., 2015; Owen et al., 2016). A possible explanation is that manual ability is related to grip strength, which is more important in men compared to women (Arnould et al., 2007; Penta et al., 2001). The construct validity of the ABILHAND-HS was further supported by our confirmation of a similar item difficulty hierarchy for QuickDASH items in our patients sample. Notably, those ABILHAND-HS activities that require a great amount of force (e.g. 'Opening a screw-topped jar') have been reported to likewise be among the most difficult items in the DASH and QuickDASH (Franchignoni et al., 2011, 2010) and in the MAM-16 (Chen et al., 2005).

The ABILHAND-HS, developed using Rasch methodology, has several advantages over questionnaires developed using classical test theory. These are summarized in Table 2.4. Firstly, the ABILHAND-HS can tolerate missing responses, which enables it to remain valid even in patients who scarcely perform some of the queried activities. Secondly, the ability to analyze response patterns can identify those patients whose responses do not fit the model due to random or careless answers, a particular injury or comorbidities. Finally, the high precision of the ABILHAND-HS items minimizes the need for interpretation, thereby allowing more reliable comparisons between patients (e.g. recreational activities involving force or impact are broken down into the items 'doing push-ups', 'practicing a racket sport').

Table 2.4. Pros and cons of the ABILHAND-HS.

	Feature	Benefit
Pros	Unidimensional and linear scale	Quantitative comparisons of manual ability can be made between patients, between treatments and along follow up
	Invariant scale calibration validated in four diagnostic groups	Unbiased comparisons of manual ability can be made within and between clinical subgroups (different HS diagnostics, stage of recovery)
	Precise item definition	Inaccurate responses and guessing are avoided
	Amenable to incomplete responses	The test is specific to activities really performed by the patient
	Capacity to analyze response patterns	The test can identify unexpected patient responses linked to patient specific behaviors, random or careless answers or comorbidities
	Precision of the measure	The standard error of measurement is specific to each patient measure allowing statistical assessment during follow-up
	Feature	**Compensation**
Cons	Yet another test for hand surgery outcomes evaluation	Takes five minutes to complete
	Complex statistical background	Necessary for the analyst, but not for routine clinical use
	Use of dedicated computer programs	A free web service (www.rehab-scales.org) can be used to interpret patient responses
	Ceiling effect of 7.9% of records	Well below the maximum recommended allowance of 15%

Limitations of this research include a sample of patients with hand and wrist disabilities from one hand surgery outpatient clinic. The unbalanced diagnostic groups and genders might have influenced the item calibrations. However, the gender distribution is similar in studies involving DRF and

CTS (Beaton et al., 2005; Levine et al., 1993; MacDermid et al., 1998) and the DIF analysis allowed us to select items where the diagnosis and gender effects were absent. Future studies with a larger sample size should confirm or refine our findings. Furthermore, the way participants responded to the 90 items in the questionnaire may not be the same as responses to the final 23-item instrument. One limitation to the availability of Rasch model-based questionnaires is that they have a complex statistical background and require the use of dedicated computer programs that are not easy to learn and implement. To facilitate and spread the use of the ABILHAND-HS, a website (www.rehab-scales.org) developed by Université catholique de Louvain and Arsalis, a spin-off of the ABILHAND authors' laboratory, can be used to convert the questionnaire raw scores into manual ability measures. The web service is free-to-use for daily practice in clinical and research applications although a license is required for commercial applications and for clinical trials.

2.5. Conclusion

ABILHAND-HS was demonstrated to be a successful adaptation for application in HS patients. The resulting scale was shown to be a valid, patient-oriented, clinically meaningful and precise instrument. It targets commonly performed manual activities and allows stable and linear measurement of manual ability over multiple time points in patients treated for DRF, BTA, CTS, or HWS. The scale reveals unexpected responses that may provide clues regarding the patient's clinical state, as summarized at www.rehab-scales.org. The questionnaire is available online, and the web service is free-to-use for daily practice in clinical and research applications although a license is required for commercial applications and for clinical trials. Future research should include more patients with HWS, as well as other diagnoses such as tendinopathies, ligamentous injuries and complex hand injuries, and an assessment of scale responsiveness.

CHAPTER 3

ABILHAND-HS: a linear scale for outcome measurement in hand surgery

Published as: El Khoury G, Penta M, Barbier O. ABILHAND-HS: a linear scale for outcome measurement in hand surgery. J Hand Surg Eur Vol. 2021 Feb 8:1753193421991485.

The ABILHAND questionnaire is a Rasch-model (Rasch, 1980) built measure of manual ability (Penta et al., 1998). It provides an invariant linear scale allowing quantitative comparisons of manual ability between patients and over time. Manual ability is defined as the capacity to manage daily activities with the upper limbs whatever the strategies involved. The adaptation of ABILHAND to hand surgery (HS) was recently reported (El Khoury et al., 2020a). Patients were asked to indicate their perceived level of difficulty associated with completing manual activities without technical or human assistance, regardless of the hand used, on a three-level scale: impossible, difficult, or easy. Successive Rasch model-based analyses led to the selection of 23 items that constitute a unidimensional manual ability scale.

We illustrate the use of the ABILHAND-HS scale in clinical practice. A 55-year-old patient with a displaced distal radial fracture of her right, dominant wrist underwent an osteosynthesis with an anterior plate. Figure 3.1 shows her manual ability evolution. The panels show her responses at two (T1), three (T2) and 12 months (T3) postoperatively. A website (www.rehab-scales.org) developed by Université catholique de Louvain and Arsalis, a spin-off of the ABILHAND authors' laboratory can be used to determine the patient's manual ability at each time-point. The website implements a Rasch analysis routine that has been validated against one of the most popular Rasch analysis software packages (RUMM2010, RUMM Laboratory Pty Ltd., Perth, Western Australia, Australia). The web service is free-to-use for daily practice in clinical and research applications although a license is required for commercial applications and clinical trials. For each assessment, the patient's manual ability is reported in logits and in centiles, with the associated 95% confidence interval (CI). Patient's improvement (0 logits (50%) at T1, 2.2 logits (68%) at T2 and 3.2 logits (77%) at T3) can be quantified on a linear scale. The CI, which reflects the precision of the instrument, is larger at the extremities of the scale (T3) compared to the center (T1), and it increases with missing responses (T1 compared to T2).

Using the web service, the clinician can verify the response pattern coherence by comparing the observed responses with the expected ones, given the patient's manual ability. Unexpected responses are those that lie outside the CI. For instance, the patient overestimated the difficulty of ''Spreading butter on a slice of bread'' at T1 and T2, and "Shuffling and dealing cards" at T2. The clinician can try to make sense of the unexpected responses. A small number of them can be found as part of a normal response pattern, but they can also be due to random answers, not following test instructions or additional comorbidities.

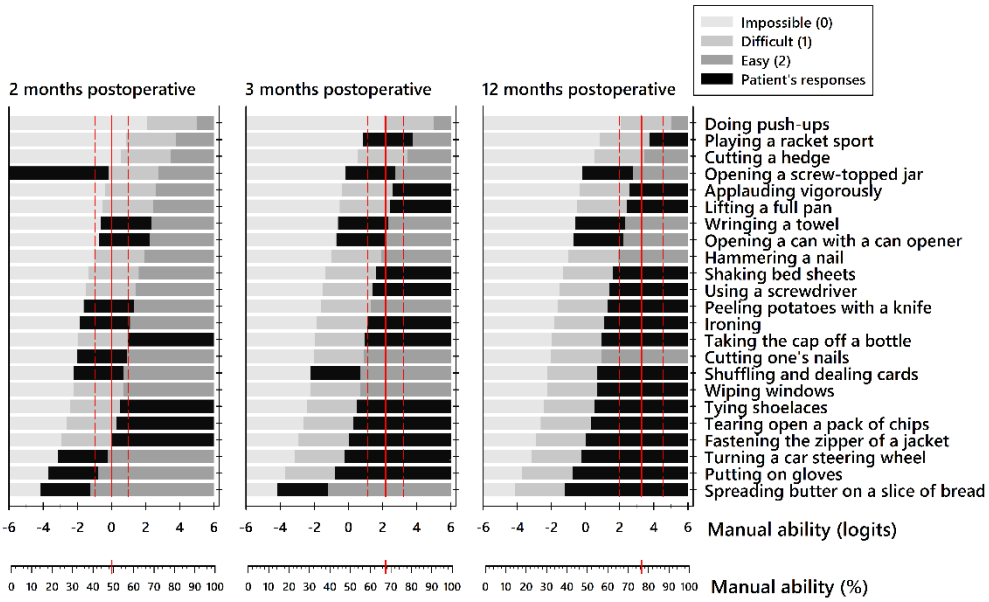

Figure 3.1. A sample scoring form showing the evaluation of a patient with a distal radial fracture of her right wrist at three time-points after surgery. Items are ordered from most (top) to least (bottom) difficult. Horizontal grey bars: patients' expected response to each item as a function of manual ability. The red vertical line represents the patient's manual ability, the dashed lines the confidence interval (+/- 1.96*standard error). The figure allows analyzing the response coherence by comparing the observed responses (black bars) to the expected responses (located inside the 95% confidence interval).

One might wonder about the advantages of this new scale, given the plethora of well-established questionnaires (El Khoury et al., 2020a). One major asset is the conversion of the ordinal raw scores into a true interval and linear unit. Raw scores represent discontinuous levels, whereas the logit is continuous and presents a fixed unit along the measurement scale. This allows for quantitative comparisons between different treatments and over time. The ability to tolerate missing responses enables the ABILHAND-HS to remain valid when some of the queried activities are rarely performed or not permitted during the recovery period; this is the most manifest at T1 (nine missing responses). The ability to analyze response patterns can single out patients' answers that are unexpected given their manual ability level. For example, a patient who answers "easy" on difficult items and "impossible" on easy items should elicit further investigation. Items are ordered by their difficulty, thus item hierarchy can be used for goal setting during the rehabilitation process.

The ABILHAND-HS was developed using the Rasch model to assess manual ability in hand surgery patients. It is oriented towards real daily life behaviors and is intended to be independent of the limb(s) or strategy used and unbiased by activities that are never performed with the affected hand or avoided during recovery. It can be used equally to measure patients with various types of impairments, as item difficulties have been shown to be stable across the tested pathologies (El Khoury et al., 2020a). This new instrument still requires validation across cultures and more hand surgical pathologies (such as tendinopathies, ligament injuries), as well as a study of its responsiveness. Nonetheless, the ABILHAND-HS is a clinically valid and methodologically sound scale for individual patient evaluation and follow-up. Together with the web-based data analysis service, it achieves high standards of functional assessment and treatment follow-up through a robust instrument for outcome measurement.

CHAPTER 4

Minimal clinically important difference and responsiveness of the ABILHAND questionnaire for hand surgery

Submitted for publication

We recently reported the adaptation of the ABILHAND scale to hand surgery (HS) patients. The purpose of the present study was to examine its responsiveness and minimal clinically important difference (MCID). Eighty-seven patients were assessed multiple times with the ABILHAND-HS questionnaire and with a global rating of change scale (GRCS). Responsiveness was tested according to both group-level and individual-level approaches. Mean score change, effect size, standardized response mean and reliable change index were calculated for groups of patients according to their GRCS. The responsiveness indices showed that the change in manual ability measures was higher in patients who reported a great improvement in their perceived status. On an individual level, the proportion of patients with a significant improvement was higher with the increase in the GRCS. A MCID of 0.50 logits (4.2 centiles) was determined based on the ROC curve, the value corresponding to a small effect size and the value of the change score in the "minimal improvement" group. The ABILHAND-HS questionnaire showed a good sensitivity to change and can thus be used for the evaluation of the effect of treatments for hand surgery and in a research setting.

4.1. Introduction

Hand surgery (HS) practice is shifting towards evidence-based treatments with the aim of providing the best results when treating patients. The growing need to assess treatment outcomes at the patient level has led to the development of patient-reported outcome measures (PROMs). Their importance has been recognized and they have been increasingly used as the primary outcome in clinical studies (Swiontkowski et al., 1999). A prerequisite for meaningful use of such PROMs is the quality of their clinimetric properties (Terwee et al., 2007).

Using the Rasch model, we recently developed and validated the ABILHAND questionnaire (Penta et al., 1998) to measure manual ability in HS patients (El Khoury et al., 2020b). The units of the scale are "logits", a probabilistic unit that defines the pass/fail probability ratio for a patient to be able to achieve an activity: the higher the logit value, the higher the probability that a patient will manage an activity easily. Logits can be converted to centiles for a more intuitive clinical interpretation. This new questionnaire presents very good psychometric properties such as linearity, unidimensionality, invariance and construct validity. Such a scale would allow for unbiased comparisons between patients and different treatment effects (Wright and Linacre, 1989). Nevertheless, its responsiveness has not been studied yet. Responsiveness, or the sensitivity to change, reflects the ability of a scale to detect a change over time when it occurs (Guyatt et al., 1989), and is a required psychometric quality for any instrument to be used in clinical studies for treatments evaluation (Terwee et al., 2007). Within this change, the minimal clinically important difference (MCID) is the minimum change in a score that indicates a meaningful change in the patient's status (Jaeschke et al., 1989).

The use of a questionnaire with known responsiveness and MCID could help quantify the effects of different treatments on manual ability. The aim

of this study was to investigate the responsiveness of the ABILHAND-HS questionnaire and to determine its MCID in a sample of HS patients.

4.2. Methods

4.2.1. Patients

Data were prospectively collected from patients recruited from the HS consultation center at Cliniques Universitaires Saint-Luc, Belgium. To be included, patients had to be over 18 years old and to read and understand French. The exclusion criteria included any comorbidity that may impede manual ability substantially (e.g. tremor, paralysis, or active rheumatologic disease) and any mental/cognitive dysfunction. Patients provided written informed consent to participate. This study was approved by the ethical committee of Cliniques Universitaires Saint-Luc-Université catholique de Louvain (N° B403201523492).

4.2.2. Procedures

Patients were asked to indicate their perceived level of difficulty associated with completing the activities without technical or human assistance, independent of the hand used to perform the activity on a three-level scale: impossible (0), difficult (1), or easy (2). Activities not attempted during the last week were treated as missing responses.

For the first evaluation, patients were given instructions and interviewed by the experimenter (GEK). For the follow-up evaluation, patients were asked to provide their subjective assessment of clinical evolution on a five-level global rating of change scale (GRCS) (Jaeschke et al., 1989) in comparison to their previous evaluation: great deterioration, minimal deterioration, no change, minimal improvement and great

improvement. To minimize ambiguity and ensure valid information, the anchor question was formulated according to Kamper et al (Kamper et al., 2009). The question reads: "With respect to your hand pathology, how would you rate your current condition compared to the last assessment?" The follow-up data were collected during a consultation or were sent by mail.

Patients presenting with the maximum score at the first assessment were excluded from the study, provided they did not judge their situation as "deteriorated", due to the ceiling effect when assessing improvement in these patients. Twenty-four patients were assessed more than twice and treated as distinct entries: one for each pair of consecutive assessments (i.e one between the first and the second assessment, and another one between the second and third assessment). Three patients did not answer the GRCS and eight presented with maximum scores at the first evaluation, these were thus excluded from the study. Our final sample consisted of 116 records from 87 patients (Table 4.1).

Table 4.1. Sample characteristics (N = 87)

Characteristic	N (%)
Gender	
Women	62 (71%)
Men	25 (29%)
Mean age (SD; range), years	62 (14.6; 26-93)
Diagnostic *group*	
Distal radius fracture (DRF)	43 (49%)
Basal thumb arthritis (BTA)	10 (11%)
Carpal tunnel syndrome (CTS)	33 (38%)
Heavy wrist surgery (HWS)	1 (1%)
Hand dominance	
Right	78 (90%)
Left	7 (8%)
Ambidextrous	2 (2%)
Involved dominant hand	
Yes	59 (68%)
No	28 (32%)
Mean (SD) time between follow-up assessments, days	181 (153)

4.2.3. Data analysis

Patients' responses to ABILHAND-HS were first converted into linear measures of manual ability (in logits and centiles) using the Rasch model, implemented with the RUMM2030 software (RUMM Laboratory Pty Ltd., Perth, Western Australia) (El Khoury et al., 2020b). The different assessments could then be treated as a continuous variable and be quantitatively compared. Standard errors of measurement (SEM) associated with the ability level of each patient were displayed by the software. Change scores

were calculated as the difference in manual ability levels between the two assessments. To make sure that the GRCS assesses the same construct measured by ABILHAND-HS under longitudinal investigation, the correlation between the GRCS and the change score had to be at least fair ($r>0.30$) (Revicki et al., 2008).

4.2.4. Responsiveness

The sensitivity to change of ABILHAND-HS was tested on group and individual levels. Groups of patients were constituted according to their response to the GRCS. Few patients reported minimal (n=8) or great (n=5) deterioration on the GRCS and were thus combined into a single "deterioration" group. The mean change (difference between the two measures) was calculated for each group.

The effect size (ES) and standardized response mean (SRM) were computed in the four groups of patients. The ES (Kazis et al., 1989) was calculated by dividing the mean of the difference between the two assessments by the standard deviation (SD) of the first measure. The value of the effect size represents the number of SDs by which the scores have changed from baseline. The standardized response mean (Liang et al., 1990) was calculated by dividing the mean change between the two assessments by the SD of the change. Higher effect sizes and standardized response means correspond to a higher magnitude of change. According to Cohen benchmarks, an ES of 0.2 is considered small, 0.5 moderate, and 0.8 large (Cohen, 1988). The same values apply for the interpretation of the SRM (Beaton et al., 1997). Responsiveness indices were expected to be larger in groups of patients who reported a larger change compared to those who reported a smaller change or a stable functional status.

The individual approach to testing the ABILHAND-HS sensitivity to change consisted in computing the reliable change index (RC) for each

patient. The RC was first proposed by Jacobson et al. (Jacobson et al., 1984) and was later modified by Christensen and Mendoza (Christensen and Mendoza, 1986). The RC is based on the standard error of measurement, and indicates to what extent the observed change exceeds the random error associated with the measuring instrument (Crosby et al., 2003).

$$RC = \frac{m_2 - m_1}{\sqrt{(SE_2)^2 + (SE_1)^2}}$$

where m_1 and m_2 are the ability measures of the first and the second evaluations, respectively, and SE_1 and SE_2 are their associated standard errors of measurement. Therefore, a value above 1.96 or below -1.96 indicates a significant improvement or deterioration, respectively (Jacobson and Truax, 1991).

4.2.5. Statistical analysis

Statistical analyses were completed in IBM SPSS Statistics for Windows, version 25 (IBM Corp., Armonk, N.Y., USA). Data normality was verified using Q-Q plots. Non-parametric tests were used when the data was found to be non-normal. A paired-samples t-test (two-tailed) was used to compare the manual ability levels at baseline and follow-up. Spearman correlation coefficient was calculated for the association between the GRCS and the manual ability change score, and between length of follow-up, the change score and the GRCS. A Mann-Whitney test was conducted to compare the differences in manual ability change between the "minimal improvement" group and the adjacent categories ("no change" and "great improvement"). The null hypothesis was rejected when the p-value was below 0.05.

4.2.6. Minimal clinically important difference

The value of the MCID was estimated using different methods:

Mean change approach. The MCID was determined as "the smallest difference in score which patients perceived as beneficial" (Juniper et al., 1994) by computing the mean change in the minimal improvement group (Sloan et al., 2003).

Small effect size. Based on a comprehensive review of the literature, Samsa et al. (Samsa et al., 1999) advocated that an effect size of 0.2 (small ES) could serve as an appropriate definition of a MCID. We verified these findings by looking at the effect size value in the "minimal improvement" group.

Standard error of measurement. The SEM takes into consideration that some observed change might be due to random error of measurement. Wyrwich et al. proposed that the one-SEM criterion could serve as a validated method for identifying the MCID (Kathleen W. Wyrwich et al., 1999; K.W. Wyrwich et al., 1999).

Receiver Operating Characteristic ROC curve. The ROC curve (Metz, 1978; Ward et al., 2000) plots sensitivity (i.e. the true positive rate) against 1 – specificity (i.e. the false positive rate). The ABILHAND-HS change score was considered true when the direction of change corresponded to the rating on the GRCS (i.e. a positive change with an "improvement" rating and a negative change with a "deterioration" rating). For the "no change" group, the change in ability measure was considered true (i.e. no change in ability measure) if the value of the RC was between -1.96 and 1.96, indicating a non-significant change. The entire cohort was used to derive the ROC curve, rather than the groups of patients adjacent to the dichotomization point, to increase precision and obtain more logical estimates of the MCID (Turner et al., 2009). The area under the curve (AUC) can be interpreted as the probability that a randomly chosen patient with a major improvement will have a higher manual ability measure than another random patient with an unimportant change (Wright et al., 2011). The greater the AUC, the greater the test is able to distinguish patients who have improved from those who

have not. AUC values between 0.7 and 0.8 are considered acceptable, values above 0.8 are considered to have excellent discrimination (Copay et al., 2007; Hosmer and Lemeshow, 2004). The optimal cutoff was chosen as the point that jointly maximized sensitivity and specificity, and thus lead to the least amount of misclassifications (Franchignoni et al., 2014).

The MCID value was estimated by integrating the results of the above methods, giving more weight to anchor-based procedures (mean change and the ROC curve).

4.3. Results

The mean ability level was 0.98 logits (56.7 centiles) at baseline and 1.75 logits (63.1) centiles at follow-up, indicating an overall increase in manual ability between consecutive assessments ($t=-4.08$, 115 df, $p<0.001$). Spearman correlation coefficient between the GRCS and the manual ability change score was 0.49 ($p<0.001$) indicating an overall coherence between the measured change in manual ability and the patients' perception of change. Length of follow-up was neither correlated to the measured change ($\varrho=0.105$, $p=0.26$), nor to the GRCS ($\varrho=-0.056$, $p=0.55$).

Patient GRCS distribution, mean change and responsiveness indices are reported in Table 4.2. The mean change score increased with the GRCS. The ES and SRM were small for the "minimal improvement" group and large for the "great improvement" group. The difference in change scores between the "minimal improvement" group and the "great improvement" group was statistically significant ($Z=-2.6$, $p=0.009$), and non significant between "minimal improvement" and "no change" ($Z=-1.2$, $p=0.22$).

Table 4.2. Responsiveness indices based on group approach

	Global rating of change scale			
	Deteriorated n=13	**No change n=22**	**Minimal improvement n=21**	**Great improvement n=60**
Mean change, logits (centiles)	-0.62 (-5.2)	-0.35 (-2.9)	0.48 (3.9)	1.58 (13.11)
Effect size	-0,36	-0,18	0,21	0,98
Standardized response mean	-0,43	-0,19	0,24	0,86

Based on the values of the reliable change (RC) index obtained for the individual approach, patient records could be divided into four categories, according to limits of significance: 1) significant improvement (RC > 1.96), 2) improvement (0 < RC < 1.96), (3) deterioration (-1.96 < RC < 0) and 4) significant deterioration (RC < -1.96). No patient had an unchanged score (RC=0). Patient proportions in each of these categories are shown in Figure 4.1. The proportion of patients with an improvement or a significant improvement was higher with the increase in the GRCS. For example, 18% of patients who reported a "minimal improvement" had a RC index indicating a significant improvement. This proportion increased to 40% in the group of patients who reported a "great improvement".

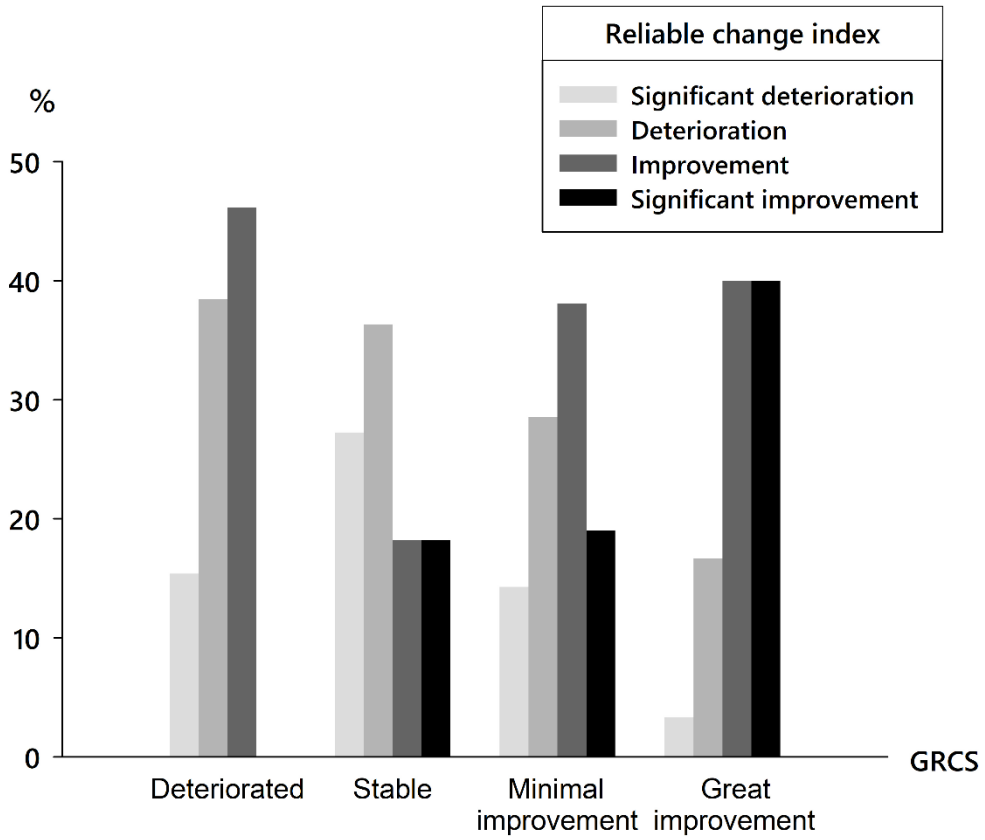

Figure 4.1. Patients distribution based on the reliable change (RC) index and the global rating of change scale (GRCS). Patients were divided into four categories based on RC significance level (significant deterioration <-1.96; deterioration -1.96-0; improvement 0-1.96; significant improvement >1.96). The proportion of patients with an improvement or a significant improvement increased with the GRCS.

MCID estimation. Based on the ROC curve (Figure 4.2), the cut-off point that best identified meaningful improvements in functional status with 75% sensitivity and 86% specificity corresponded to 0.51 logits (4.2 centiles). The AUC was 0.815 (95% CI: 0.73, 0.90), which corresponds to excellent discrimination. When compared to the mean change in the minimal improvement group of 0.48 logits (3.9 centiles), which corresponded to a small effect size of 0.21, and to the median SEM associated with patients' ability measures of 0.52 logits (4.3 centiles), the three estimates of MCID were

quite close, with an average of 0.50 logits (4.2 centiles). Among the patients that identified themselves as "improved", the value of 0.50 logits (4.2 centiles) correctly identified 77% of the patients.

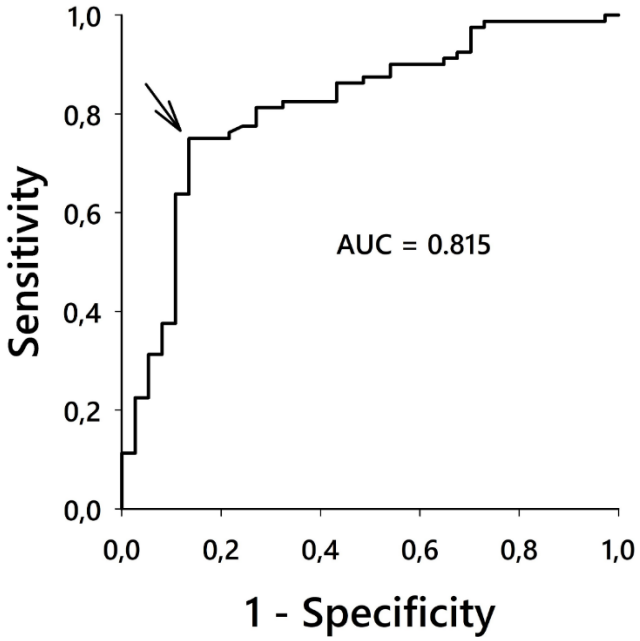

Figure 4.2. Receiver-operating-characteristic (ROC) curve. The ROC curve shows the accuracy in identifying patients with a minimal improvement compared to no improvement. The arrow shows the value that maximizes sensitivity and specificity, corresponding to a change score of 0.51 logits (4.2 centiles). AUC: area under the curve.

4.4. Discussion

The responsiveness of the ABILHAND-HS questionnaire was investigated in 116 entries from 87 patients by computing responsiveness indices after separation into four groups (deterioration, no change, minimal improvement and great improvement) based on their GRCS. The mean change in manual ability measures, ES and SRM increased with the GRCS. Although we obtained a clear hierarchy and a good separation between

categories confirming our initial hypothesis, the difference between "no change" and "minimal improvement" was non-significant. However, the statistical significance depends not only on the magnitude of the change, but also on the sample size and the variability of the measure, and conveys little information about the clinical meaningfulness of that change (Crosby et al., 2003). The statistical tests performed in our sample were thus underpowered, mainly because of a large variance and our limited sample size. The responsiveness of the ABILHAND-HS was also investigated according to an individual approach using the RC index. The proportion of patients with a significant improvement in manual ability level increased with the GRCS, which is consistent with the results of the group-level approach (increase in ES and SRM). Consequently, the individual-level approach provides clinicians an alternative method of drawing conclusions from group results to individuals.

The MCID was estimated by using four methods (mean change, small ES, one-SEM and ROC curve) that yielded approximately the same value of 0.50 logits (4.2 centiles). The ABILHAND MCID was found to be equal to 0.47 logits for rheumatoid arthritis patients (Batcho et al., 2011), and 0.26 to 0.35 logits in patients with stroke (T. Wang et al., 2011). These minor differences with our MCID estimation (0.50 logits) can be attributed to patient characteristics and different methods of estimation. While each of these methods presents inherent limitations, the convergence of the MCID computed with distribution- and anchor-based methods reinforces our confidence in the MCID estimation. Mean change is a poor descriptor of non-normally distributed data, which is sometimes the case in clinical change, and is susceptible to outliers. The SEM is not constant across the range of ability (it is the largest at both extremes of the scale). The ES is influenced by the sample distribution at baseline (i.e. for the same given change score, a larger baseline SD will give a smaller resultant ES). The ROC approach accommodates skewed data, uses all available data, is not vulnerable to a small number of values within a category, and maximizes the number of

individuals correctly classified (Turner et al., 2009). The chosen MCID value of 0.50 logits (4.2 centiles) separated the GRCS categories "no change" and "minimal improvement" with 75% sensitivity and 86% specificity. These results are comparable to the results for the DASH (82% sensitivity; 74% specificity; 79% correctly classified) and the QuickDASH (79% sensitivity; 75% specificity; 78% correctly classified) (Franchignoni et al., 2014). The obtained values of ES and SRM were in line with the values obtained in other studies for the DASH, QuickDASH, Carpal Tunnel Questionnaire and Michigan Hand Questionnaire (Chatterjee and Price, 2009; Hong et al., 2018; Kotsis and Chung, 2005; da Silva et al., 2020).

There is no standard format for the GRCS question, thus the wording of the GRCS question has the potential to influence patients' responses, and hence the MCID estimation (Sloan et al., 2003). Different authors have used different cutoffs for determining the minimum change. For instance, The MCID has previously been derived using small (e.g., 1-3 on a -7 to +7 point GRCS (Wyrwich and Wolinsky, 2000)) or moderate change (e.g., 4-5 on -7 to +7 point GRCS (Cleland et al., 2008)). To date, there is no consensus on the optimal threshold to use (Turner et al., 2010) and this threshold is often arbitrary (Copay et al., 2007). We used a 5-level GRCS, which contains fewer categories than most reports. This presents the advantage of better discrimination between categories and not choosing an arbitrary cut-off to determine the MCID. The variability obtained within each category (see Fig 1) suggests that adding more categories to the GRCS would have generated more noise in the data. The GRCS may be influenced by recall bias (Schwarz and Sudman, 1994), especially over the long term, or the lack of patient ability to understand the context of improvement (Kamper et al., 2009). Although follow-up duration was variable in our study, it was not correlated with the GRCS nor with the change score. The GRCS may also be disproportionately affected by the current health status rather than the change over time (Norman et al., 1997). The GRCS may also be influenced by other domains of improvements or deterioration (such as pain), while

ABILHAND-HS measures the patient's activity (El Khoury et al., 2020b). Nonetheless, the GRCS correlated well with the measured change (ϱ =0.49), which suggests that they measure the same construct (Revicki et al., 2008).

MCID estimation values were around 0.50 logits (4.2 centiles) in our sample. We recommend using this value on a group level, as when assessing large groups of patients to compare different treatments. One must keep in mind that MCID values are dependent on sample characteristics, pathologies, and time interval between evaluations (Y.-C. Wang et al., 2011). Stucki et al. cautioned against using one general benchmark for ordinal scales, as numerically equal gains of ability will be different depending on the baseline health status (Stucki et al., 1996). Total scores obtained by adding up the values of each response are ordinal and not necessarily linear, which means that the measurement unit is not constant throughout the measurement range. The same distance between scores (e.g. from 0 to 1 and from 1 to 2) may not reflect the same amount of increase in ability. This distortion of the score is especially noticeable at the extremes of the score range, compared to the center of the scale. For example, a 2-point difference at the center may represent a smaller true score difference than a 2-point difference at the extremes (DeVellis, 2006). However, ABILHAND-HS is an interval scale and a change of 0.50 logits (4.2 centiles) corresponds to the same in manual ability whatever the initial patient ability. Nonetheless, this result should be confirmed in different samples (e.g. different upper limb pathologies) with varying baseline status.

The magnitude of change necessary to be considered meaningful may be different between the group and the individual approaches (Beaton et al., 2001; Cella et al., 2002). Relatively modest improvements at the individual level may be considered clinically important when considered at the group level (Crosby et al., 2003). For this reason, the RC index, derived from the confidence interval around the patient's ability (+/- 1.96*SEM) can be used as a guidance for the MCID of individual patients. A RC index above 1.96 or below -1.96 means that the observed change exceeds the random error

associated with the measuring instrument. When an improvement or deterioration is not reflected by the measured change, other dimensions (such as pain or psychological state) that might explain the discrepancy between the patient's assessment and the questionnaire should be assessed.

In conclusion, this study shows that the ABILHAND-HS is a responsive tool to assess the effects of different treatments in hand surgery. It can thus be used for clinical evaluation or as an outcome measure in clinical studies. Future research should aim at confirming our initial results in a larger sample and more varied diagnoses.

CHAPTER 5

Recognizing manual activities using wearable inertial measurement units: clinical application for outcome measurement

Published as: El Khoury G, Penta M, Barbier O, Libouton X, Thonnard JL, Lefèvre P. Recognizing Manual Activities Using Wearable Inertial Measurement Units: Clinical Application for Outcome Measurement. Sensors (Basel). 2021 May 7;21(9):3245. doi: 10.3390/s21093245.

The ability to monitor activities of daily living in the natural environments of patients could become a valuable tool for various clinical applications. In this paper, we show that a simple algorithm is capable of classifying manual activities of daily living (ADL) into categories using data from wrist- and finger-worn sensors. Six participants without pathology of the upper limb performed 14 ADL. Gyroscope signals were used to analyze the angular velocity pattern for each activity. The elaboration of the algorithm was based on the examination of the activity at the different levels (hand, fingers and wrist) and the relationship between them for the duration of the activity. A leave-one-out cross-validation was used to validate our algorithm. The algorithm allowed the classification of manual activities into five different categories through three consecutive steps, based on hands ratio (i.e., activity of one or both hands) and fingers-to-wrist ratio (i.e., finger movement independently of the wrist). On average, the algorithm made the correct classification in 87.4% of cases. The proposed algorithm has a high overall accuracy, yet its computational complexity is very low as it involves only averages and ratios.

5.1. Introduction

Hands can be affected in different neurologic, rheumatologic, degenerative or traumatic conditions. To evaluate this manual impairment, physicians rely on medical history and clinical examination, but have also several tools at their disposal. For instance, they can use diagnostic tests such as electromyography and patient-reported outcome measures that reflect the patient's point of view (Barbier et al., 2003). Motion capture analysis can also provide additional information, though it is more commonly used in research rather than in a routine clinical setting. Medical practice has shifted towards evidence-based treatments with the aim of providing the best results when treating patients. Therefore, robust outcome evaluations are needed to assess the effectiveness and reliability of a treatment (Porter, 2009).

An activity is defined in the International Classification of Functioning, Disability and Health (ICF) as the execution of a task or action by an individual (World Health Organization, 2001). Measuring the activity domain is a key point in determining the impact of different treatments on functional recovery, as the consequences of a pathology on patients' functioning are the most manifest through their inability to carry out activities of daily living (ADL) (Arnould et al., 2007). Activity performance cannot be measured directly, but can either be inferred by direct observation, which is time consuming in practice, or can be self-reported by patients through questionnaires.

Questionnaires can provide self-reported measures focused on the patients' perceptions of their activity limitations. They inform clinicians on how well patients manage their activity in their home environment. For example, ABILHAND is a questionnaire that measures manual ability through activities that present a common perceived difficulty among patients (Penta et al., 1998). It provides an invariant linear scale allowing quantitative comparisons of manual ability between patients and over time. The units of this scale are expressed in logits, and can be converted into

centiles for a more intuitive clinical interpretation. The scale has been validated in populations with various pathologies (Arnould et al., 2004; Durez et al., 2007; El Khoury et al., 2020b; Penta et al., 2001; Vandervelde et al., 2010; Vanthuyne et al., 2009). Other questionnaires such as the Disabilities of the Arm, Shoulder and Hand (DASH) (Hudak et al., 1996), the Patient-Rated Wrist Evaluation (PRWE) (MacDermid et al., 1998) and the Carpal Tunnel Questionnaire (CTQ) (Levine et al., 1993) have been developed to measure different aspects of upper limb function. These self-reported measures are based on the respondent's memory of the perceived difficulty and their ability to accurately judge their capability (Holsbeeke et al., 2009). Items that compose these questionnaires are representative of the patients' daily manual activities (e.g., using a spoon or tying shoelaces).

Another complementary approach to that of the questionnaires would be a direct assessment of the patient's actual activities. A direct assessment could be used to monitor a patient's actual activity objectively, without relying on the patient's memory, and systematically, witnessing what activities the patient actually does or does not do. The ability to monitor activities of daily living in the patient's natural environment could become a valuable tool for clinical decision-making, evaluating healthcare interventions, and supporting and tracking rehabilitation progress. Inertial sensors have been used for monitoring activities as they are small, affordable, and generally unobtrusive (Yang and Hsu, 2010). They have been used for upper limb motion analysis with good accuracy and reliability (Cuesta-Vargas et al., 2010; Zhou et al., 2008). They have been shown to be useful for clinical applications (Thanawattano et al., 2015), and proved to be more sensitive than questionnaires to detect changes in shoulder movement, thus adding a complementary objective component to outcome measurement (Körver et al., 2014).

Different authors have worked on recognizing upper limb movements using accelerometry alone (Biswas et al., 2014; Lemmens et al., 2015) or in combination with surface electromyography (Roy et al., 2009), and on

building devices that could track hand use (Rowe et al., 2013). For instance, the "manumeter" determines hand use by tracking the total angular distance traveled by the wrist and fingers using magnetometers (Rowe et al., 2013). This device is able to track global hand use, but performs poorly for tasks requiring small yet intensive movements such as handwriting (Rowe et al., 2014). Another limitation is the interaction with ferromagnetic objects, which are commonly used in everyday life, and can alter the device readings.

As a complementary approach to questionnaires that are common to a patient population, a hand activity monitoring device that not only tracks the global hand use, but is also able to categorize the manual activities that are actually performed, would offer a more personalized approach and would have implications in many aspects of patient care. In this paper, we show that a simple algorithm is capable of classifying manual activities of daily living using data from wrist- and finger-worn sensors.

5.2. Materials and Methods

5.2.1. Prototype

We used a prototype device (InSense©, Arsalis, Belgium) to capture human activity signals using inertial measurement units (IMUs). The device is shown in Figure 5.1A. Each sensor integrates a triaxial accelerometer and a triaxial gyroscope. The measurement range is ±16 g and ±2000 °/s for each axis of accelerometer and gyroscope, respectively. The device is wired and transmits sampled sensor data to a laptop computer via a USB interface. The sensors are small in size (9.4 × 8 × 5.5 mm) and lightweight enough (2 g) to be worn comfortably without altering the hand movements. The inertial signals of all sensors are sampled synchronously (inter sensor delay < 0.125 ms) with a 16-bit resolution at a rate of 500 Hz.

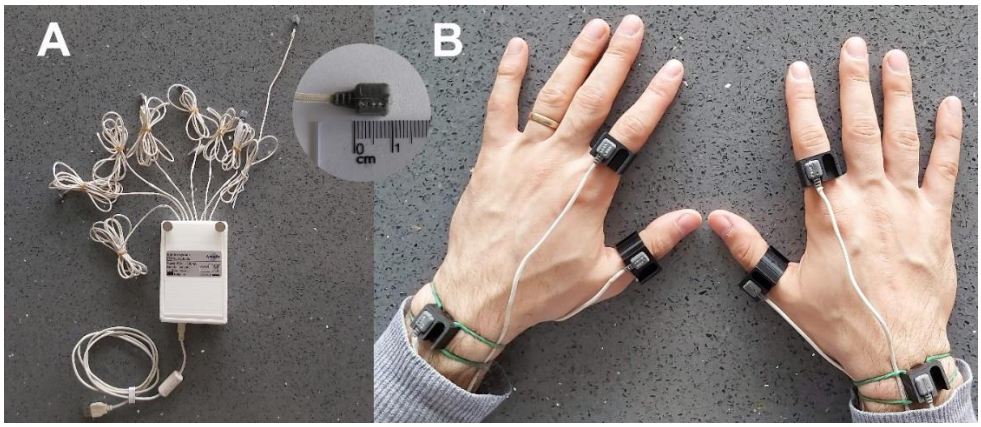

Figure 5.1. (A). Photograph showing the device prototype, which consists of eight inertial measurement units connected to a processor. A close-up of one of the sensors is shown. (B). Photograph showing the placement of the sensors on 3D-printed supports on the participant's hands.

5.2.2. Sensor Calibration

Accelerometers and gyroscopes were calibrated prior to performing the experiments so that the readings were accurate and reliable. Accelerometers were calibrated by applying 0 g, 1 g and −1 g on each accelerometer of each sensor. Their calibration reported an average absolute error of 0.18% of full scale (FS) on any axis of any sensor (range: 0.05 to 0.62 %FS). Gyroscopes were calibrated using a rotating device equipped with a 1024 point resolution optical encoder that was used to determine the reference angular speed (Video S1 in Supplementary Materials). They were calibrated at angular speeds ranging from -600 to +600 °/s and reported an average absolute error of 0.23 %FS on any axis of any sensor (range: 0.13 to 0.71 %FS). Raw data were converted to physical values of angular velocity and acceleration expressed in °/s and g, respectively, using individual sensor calibration coefficients.

5.2.3. Participants

Six healthy adults participated in this study; their characteristics are detailed in Table 5.1. Participants were included in the study if they were above 18 years old and had no pathology that could affect the use of their upper limbs. The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the ethics committee of Cliniques Universitaires Saint-Luc Université catholique de Louvain (2015/26JAN/025, N° B403201523492). Participants provided written informed consent to make use of their anonymized data.

Table 5.1. Participants' characteristics.

	Age	**Sex**	**Height (cm)**	**Weight (kg)**	**Work**
Participant 1	31	F	152	42	Office worker
Participant 2	65	M	162	80	Dentist
Participant 3	28	M	173	74	Office worker
Participant 4	24	F	176	78	Student
Participant 5	31	M	171	70	Office worker
Participant 6	57	F	164	53	Housewife

5.2.4. Activities Selection

In order to explore the wide range of hand movements, activities were selected from the different pathology-specific versions of the ABILHAND questionnaire (hand surgery, stroke and rheumatoid arthritis). Items from this questionnaire have been rigorously selected to report patient-perceived difficulty unbiased by patient demographics (e.g., age, gender) nor clinical conditions (e.g., side affected, manual ability). Twelve activities were

selected to cover the whole range of measurement of the ABILHAND scale. Two additional items were added for their relevance in everyday life, namely "typing on a computer keyboard" and "using a spoon". The final list of 14 activities is shown in Table 5.2. We hypothesized that these activities could be classified into five different categories, based on the way they are actually executed. Some activities are unimanual while others are bimanual. Bimanual activities could require the action of a stabilizing hand or involve both hands equally. In addition, some activities require the use of the fingers (the fingers move independently of the wrist), while others involve the whole hand (the fingers move together with the wrist), for example, when manipulating a tool. When some manual activities could be performed in different ways (e.g., some participants brushed their hair with both hands while others used only their dominant hand), the experimenter's judgement was used to classify each activity into a category, depending on the way it was executed by the participant.

Table 5.2. List of manual activities and their respective categories.

	Activity	Category
1	Using a spoon	Unimanual
2	Drinking a cup of water	
3	Brushing one's hair	
4	Writing a sentence	Bimanual with a stabilizing hand and finger activity of the active hand
5	Spreading butter on a slice of bread	Bimanual with a stabilizing hand and global activity of the active hand
6	Opening a can with a can opener	
7	Typing on a computer keyboard	Bimanual with finger activity of both hands
8	Shuffling and dealing cards	
9	Peeling potatoes with a knife	
10	Buttoning a shirt	
11	Tying shoelaces	
12	Opening a screw-topped jar	Bimanual with a global activity of both hands
13	Lifting a full pan	
14	Wringing a towel	

5.2.5. Experimental Setup and Recordings

Participants were equipped with the prototype device sensors on the first phalanges of the first two fingers of both hands and on the wrists (Figure 1B). Sensors were fitted on 3D-printed supports in the shape of rings for the fingers and wristbands for the wrists. These sites were chosen to correspond to sites where everyday accessories are worn (watch and rings) and do not hinder activities of daily living.

Participants were asked to perform the 14 activities in a random order for five repetitions each, while sitting on a chair at a table. The tools used (e.g., can opener, pen) were from the participants' home environment. They

were instructed to perform each activity as they would do in their normal life, with no constraints except for the duration of each activity (no more than 25 seconds per repetition). Each activity started and ended with the hands still on the table, separated by five seconds of inactivity. Experiments were performed under the supervision of the experimenter.

5.2.6. Data Analysis

Each recording was processed to isolate the activity period (i.e., when the participant is actually executing the task) from inactivity periods (between two consecutive repetitions). The main goal was to focus on activity recognition, based on the assumption that the start and end of an activity were known.

Gyroscope signals were used to analyze the angular velocity pattern for each activity, as they demonstrated the most distinctive pattern compared to the accelerometers. No filter was applied to the raw data. For each gyroscope signal, the norm of the angular velocity vector was computed by combining the x, y and z components. Signals were combined to compute the hand signal (mean of the three IMUs on one hand) and the fingers signal (mean of the two IMUs placed on the fingers) for both limbs. The elaboration of the algorithm was based on the examination of the activity at the different levels (hand, fingers and wrist) and the relationship between them for the duration of the activity.

The hands ratio (HR) was calculated by dividing the angular velocity of the most active hand by that of the least active one.

$$HR = \frac{Most\ active\ hand}{Least\ active\ hand}$$

It was chosen as a criterion to differentiate between bimanual activities involving both hands equally and those involving a stabilizing hand. When both hands are involved equally, the HR is expected to be close to one.

During unimanual activities or when one hand stabilizes an object, one hand is less active than the hand performing the movement, and the HR is expected to increase.

The fingers-to-wrist ratio (FWR) was computed by dividing the fingers' angular velocity (mean of both fingers) by the wrist angular velocity.

$$FWR = \frac{Fingers' \ angular \ velocity}{Wrist \ angular \ velocity}$$

When the hand moves as a whole (e.g., when manipulating a hammer), the angular velocity in the fingers is close to that of the wrist; hence, the FWR is close to one. If the fingers are involved in independent movements (e.g., when writing), the FWR increases. The FWR was computed on the dominant hand for bimanual activities with a stabilizing hand, and by taking the average of the two hands for bimanual activities.

5.2.7. Determining Cutoff Points

The Receiver Operating Characteristic (ROC) curve was used to determine the cutoff points for HR and FWR that best discriminate between the different categories of activities (Metz, 1978). The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve plots sensitivity (i.e., the true positive rate) against 1–specificity (i.e., the false positive rate) at various threshold settings. The optimal cutoff value (i.e., threshold) was chosen as the point that jointly maximized sensitivity and specificity, hence leading to the least number of misclassifications. The area under the curve (AUC) is the measure of the ability of the classifier to distinguish between classes. The greater the AUC, the better the criterion is able to distinguish between the different categories. AUC values between 0.7 and 0.8 are considered acceptable, and values above 0.8 are considered to have excellent discrimination levels (Hosmer and Lemeshow, 2004).

5.2.8. Algorithm Validation

A leave-one-out cross-validation was used to validate our algorithm (Halilaj et al., 2018), as detailed in Figure 5.2. For each iteration of the validation, one participant was left out of the training sample, and data from the five remaining participants were used to derive cut-off values for the HR and FWR and establish the algorithm. The latter was then applied to the data of the participant left out to evaluate the performance of the algorithm. Each individual repetition was categorized using these cutoffs and following the steps laid by the algorithm. This process was repeated six times in total to compute the validation errors. Activities were considered correctly identified into their respective categories if both criteria (HR and FWR) for this category were in the right range at each step of the algorithm. The performance of the algorithm was then calculated by comparing the activity category as established by the experimenter and the categorization provided by the algorithm.

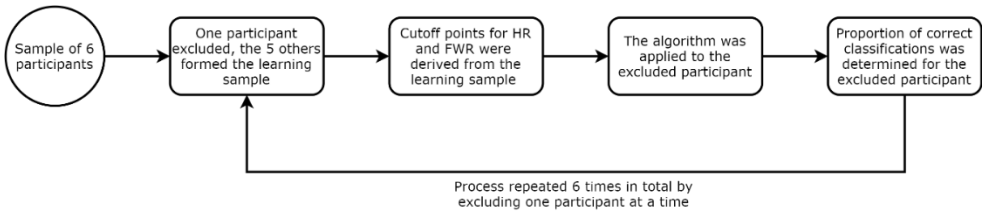

Figure 5.2. Diagram showing the validation process of the algorithm.

5.3. Results

5.3.1. Cutoff Points

The ROC curves used to determine the cutoff points for the hands ratio (HR) and fingers-to-wrist ratio (FWR) for the whole sample are shown in Figure 5.3. The sensitivity ranged between 96% and 100%, and the specificity from 85.7% to 98.7%. The AUCs for all criteria were above 0.978, providing excellent discrimination. The individual values for the different iterations of the leave-one-out cross-validation are detailed in Table 5.3. These did not vary substantially in comparison with values for the whole sample, demonstrating the robustness of the approach.

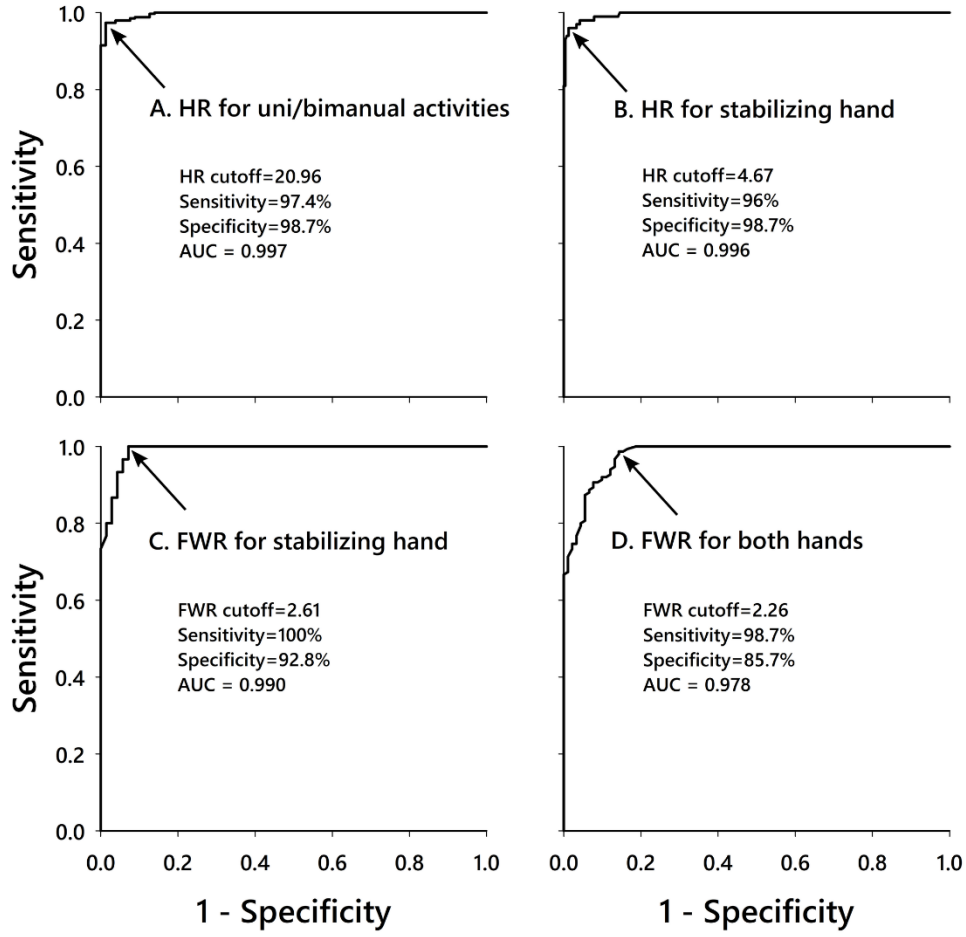

Figure 5.3. Receiver-operating-characteristic curve showing the cut-off points for the hands ratio (HR) and the fingers-to-wrist ratio (FWR). The arrows show the point that maximizes sensitivity and specificity. Panel A: HR for the discrimination between uni- and bimanual activities. Panel B: HR for the discrimination between activities involving a stabilizing hand and those involving both hands. Panel C: FWR for identifying finger activity in activities involving a stabilizing hand. Panel D: FWR for identifying finger activity in activities involving both hands. AUC: Area Under the Curve.

Table 5.3. Cut-off values for the hands ratio and the fingers-to-wrist ratio.

	HR[1] for classification between uni- and bimanual activities	HR[1] for classification between bimanual activities involving a stabilizing hand and those using both hands	FWR[1] for fingers involvement of bimanual activities using a stabilizing hand	FWR[1] for fingers involvement of bimanual activities using both hands
Participant 1 excluded	20.96	4.25	2.68	2.50
Participant 2 excluded	20.96	4.67	2.61	2.42
Participant 3 excluded	20.96	4.67	2.61	2.26
Participant 4 excluded	22.01	4.71	2.56	2.26
Participant 5 excluded	20.96	4.67	2.61	2.42
Participant 6 excluded	20.95	4.62	2.61	2.25
Whole sample	20.96	4.67	2.61	2.26

[1] HR: Hands Ratio; FWR: Fingers-to-Wrist Ratio.

5.3.2. Description of the Algorithm

The algorithm (Figure 5.4) allows the classification of manual activities into five different categories through three different steps, based on HR and FWR. The first step of the algorithm separates unimanual from bimanual activities based on HR. A HR greater than 20.96 is indicative of unimanual activities, i.e., one of the hands is over 20 times more active than the other hand. For the second step of the algorithm, a cutoff HR of 4.67 can be used to separate bimanual activities that use a stabilizing hand from those that involve both hands equally. The HR for this second step is smaller than that of the first step, as the stabilizing hand still performs low amplitude

movements. The third step separates activities based on whether the movement involves the fingers or not. A FWR larger than 2 (actually 2.61 for bimanual activities involving a stabilizing hand and 2.26 for bimanual activities involving both hands equally) means that the fingers are about two times more active than the wrists, indicating that the fingers are mainly performing the movement such as when writing or buttoning a shirt. On the contrary, "spreading butter on a slice of bread", for example, involves using the hand as a whole when manipulating a tool (global hand activity, FWR< 2). In summary, the algorithm can classify manual activities based on the involvement of the hands relative to one another, and the presence or absence of finger activity.

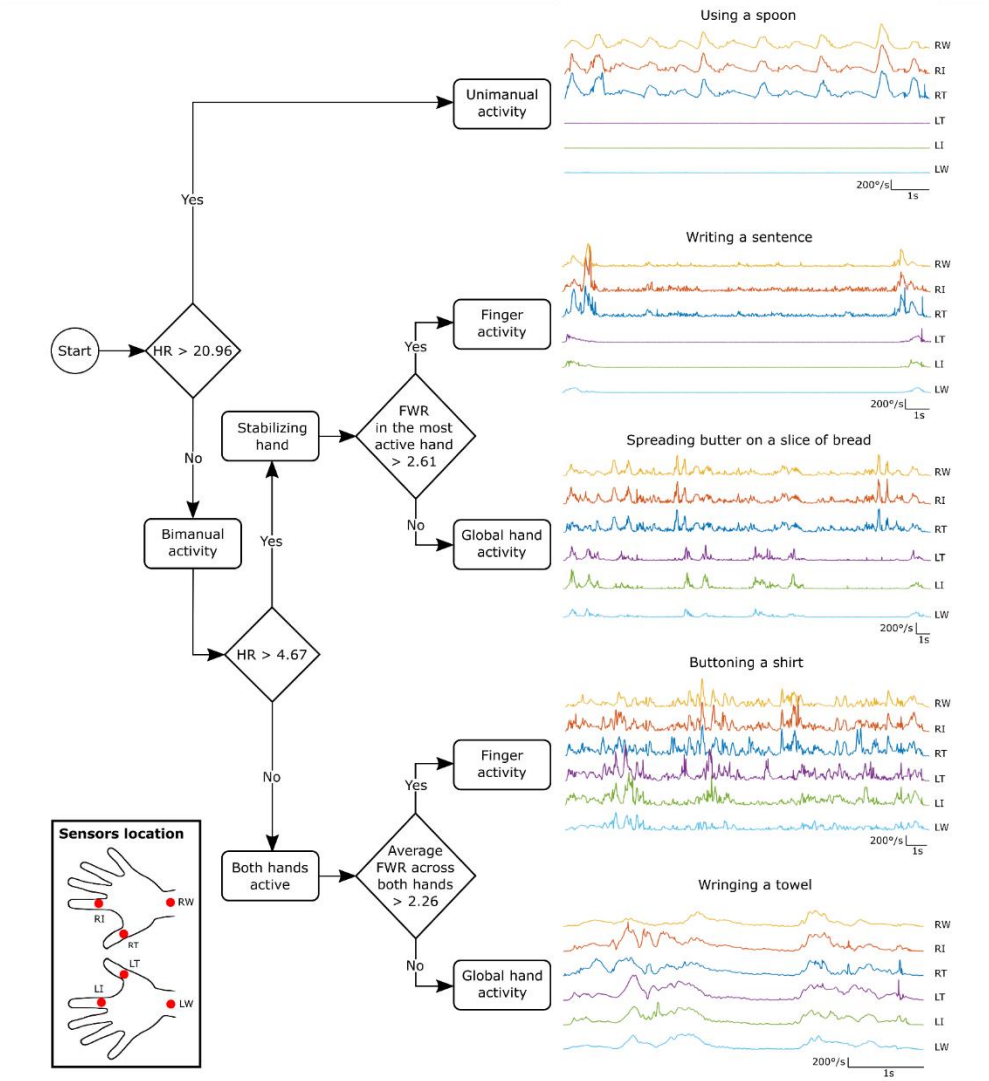

Figure 5.4. Algorithm for the classification of manual activities. The values shown for the hands ratio (HR) and fingers-to-wrist ratio (FWR) are those extracted from the whole sample. Typical traces for five tasks performed by a right-handed subject show the signals of the six sensors for one repetition of one selected activity for each category. RW: Right Wrist, RI: Right Index; RT: Right Thumb; LT: Left Thumb, LI: Left Index; LW: Left Wrist.

5.3.3. Performance of the Algorithm

Cutoff values for HR and FWR were derived from the learning sample and then tested for validation on the participant left out using the classification algorithm. An example of the validation method is shown in Table 4. For each repetition of each activity, the HR and FWR were extracted for participant 3. Each one of these values was then compared to the cutoffs derived when excluding participant 3 (see Table 5.3), according to the steps previously detailed in the algorithm. When the observed value was in the expected range, the cell was colored in green. When outside the range, it was colored in red. For example, the HR for the first repetition of "Using a spoon" was 56.73, which is >20.96 and, thus, verified the criteria for being a unimanual activity. The FWR for the fifth repetition of "Opening a screw-topped jar" was 2.43, which was slightly above the expected value for a bimanual activity with global activity of both hands (the FWR should be <2.26). The observed value was outside the range, and the cell was colored in red. This process was repeated for each one of the six participants, and the sum of correct classifications was computed.

Table 5.4. Validation of the algorithm for participant 3[1]

				Step 1					Step 2						Step 3				
				HR [2] for Uni/Bimanual Activities					HR [2] for Stabilizing Hand/both Hands Active						FWR [2] for Finger Activity/Global Hand Activity				
				Cutoff	Rep1 [2]	Rep2	Rep3	Rep4	Rep 5	Cutoff	Rep1	Rep2	Rep3	Rep4	Rep 5	Cutoff	Rep1	Rep2	Rep3	Rep4	Rep 5
Unimanual activities			Brushing one's hair	HR > 20.96	32.11	54.26	66.41	68.41	83.61	N/A						N/A					
			Using a spoon		56.73	32.44	41.81	67.44	59.21												
			Drinking a cup of water		48.90	58.22	68.75	68.21	85.08												
Bimanual activities	Stabilizing hand	Finger activity	Writing a sentence	HR < 20.96	27.72	25.11	26.97	20.13	29.52	HR > 4.67	27.72	25.11	26.97	27.54	29.52	FWR > 2.61	2.80	2.72	2.80	2.83	2.68
		Global hand movement	Spreading butter on a slice of bread		7.63	8.10	11.44	10.69	16.35		7.63	8.10	11.44	10.69	16.35	FWR < 2.61	1.72	1.62	1.61	1.61	1.49
			Opening a can with a can opener		5.92	8.52	9.98	11.93	11.41		5.92	8.52	9.98	11.93	11.41		2.29	2.67	2.01	2.10	2.10
	Both hands active	Finger activity	Typing on a computer keyboard		3.21	2.50	2.75	3.11	2.91	HR < 4.67	3.21	2.50	2.75	3.11	2.91	FWR > 2.26	5.77	5.58	5.20	5.29	5.36
			Shuffling and dealing cards		2.80	2.44	2.58	2.59	2.58		2.80	2.44	2.58	2.59	2.58		2.97	3.28	2.93	3.23	3.24
			Peeling potatoes with a knife		1.74	1.97	1.93	1.98	1.94		1.74	1.97	1.93	1.98	1.94		2.80	2.62	2.90	3.07	3.02
			Buttoning a shirt		2.20	2.01	2.42	2.16	2.67		2.20	2.01	2.42	2.16	2.67		3.00	3.13	2.88	3.27	3.16
			Tying shoelaces		2.71	4.27	2.42	3.16	2.61		2.71	4.27	2.42	3.16	2.61		2.75	2.64	2.85	2.60	2.54
		Global hand movement	Opening a screw-topped jar		1.93	1.77	1.78	2.15	1.92		1.93	1.77	1.78	2.15	1.92		2.02	1.98	2.18	1.94	2.43
			Lifting a full pan		1.51	1.52	1.54	1.50	1.51		1.51	1.52	1.54	1.50	1.51	FWR < 2.26	1.70	1.55	1.56	1.50	1.49
			Wringing a towel		2.29	2.17	2.15	2.54	2.38		2.29	2.17	2.15	2.54	2.38		2.01	1.72	2.19	1.87	1.73

[1] Green and red cells: when the value of the HR or FWR verifies the condition or not, respectively. [2] HR Hands Ratio; FWR: Fingers-to-Wrist Ratio; Rep: repetition

The performance of the algorithm on the validation sample is detailed in Table 5.5. Each column in the table represents a step in the algorithm, and the percentage of correct classification is detailed for the classification of each activity in the correct category.

For the first step, the algorithm was able to classify uni- from bimanual activities based on HR with an average accuracy of 97%. The activity "writing a sentence" was incorrectly classified in 33% of the cases as a unimanual activity. This is explained by the fact that the stabilizing hand is only active at the beginning and the end of the movement, and thus, has little influence on HR, especially as the activity lasts longer. This misclassification originated almost exclusively from two subjects (nine out of ten incorrect classifications).

For the second step of the algorithm, activities requiring a stabilizing hand were classified correctly in 95% of cases and those involving both hands equally in 98% of cases. The third step correctly identified the presence or the absence of fingers' involvement in 89 to 100% of cases, per category.

For an activity to be classified in the correct category, it had to verify the HR and FWR criteria for every repetition. On average, this was achieved in 87.4% of the activities, as shown by the overall accuracy in the last column of Table 5.5.

Table 5.5. Performance of the algorithm

				Step 1 HR [1] for Uni/Bimanual Activities		Step 2 HR [1] for Stabilizing Hand/both Hands Active		Step 3 FWR [1] for Finger Activity/Global Hand Activity		Overall Accuracy	
				Accuracy per Activity	Accuracy per Category	Accuracy per Activity	Accuracy per Category	Accuracy per Activity	Accuracy per Category	Accuracy per Activity	Accuracy per Category
Unimanual activities			Brushing one's hair (19) [2]	100%	97%	N/A	N/A	N/A	N/A	100%	97%
			Using a spoon	100%						100%	
			Drinking a cup of water	93%						93%	
Bimanual activities	Stabilizing hand	Finger activity	Writing a sentence	67%	97%	100%	95%	100%	100%	67%	67%
		Global hand movement	Spreading butter on a slice of bread	100%		97%		100%	90%	97%	84%
			Opening a screw-topped jar (10) [2]	100%		90%		40%		40%	
			Opening a can with a can opener	100%		90%		97%		87%	
	Both hands active	Finger activity	Typing on a computer keyboard	100%		100%	98%	90%	89%	90%	86%
			Shuffling and dealing cards	100%		90%		100%		90%	
			Peeling potatoes with a knife	100%		100%		93%		93%	
			Buttoning a shirt	100%		97%		87%		83%	
			Tying shoelaces	100%		97%		73%		73%	
		Global hand movement	Opening a screw-topped jar (20) [2]	100%		100%		75%	90%	75%	90%
			Lifting a full pan	100%		100%		100%		100%	
			Wringing a towel	100%		97%		87%		87%	
			Brushing one's hair (11) [2]	100%		100%		100%		100%	

[1] HR Hands Ratio; FWR: Fingers-to-Wrist Ratio [2] When participants performed the activity differently, the number between brackets indicates the number of repetitions in the current category.

5.4. Discussion

In this paper, we show the applicability of a very simple algorithm for the categorization of 14 manual ADL. Using gyroscope data from six IMUs located on the thumb, index finger and wrist of both hands, we were able to classify manual ADL into five categories. The proposed algorithm has a high overall accuracy, yet its computational complexity is very low as it involves only averages and ratios of sensor measurements.

Our algorithm was able to classify manual activities into their correct category in 87.4% of cases. The poorest performance in categorization corresponded to the activity "writing a sentence" (category "bimanual activities with a stabilizing hand and finger activity of the active hand"), for which the accuracy was 67% on average. Our results show that it is a borderline activity that can be performed using only one hand if the support is stable enough. Contrary to lower limb movements, most manual activities are complex to analyze, mainly because they are non-cyclical and variable (Rau et al., 2000). Differences in movement patterns exist across individuals and across repetitions by the same individual. This was especially evident in our study for the activity "brushing one's hair". Participants used either one or both hands to brush their hair, and the activity was, thus, considered either unimanual or bimanual depending on the actual performance. In practice, this misclassification can be tolerated and only highlights the variability across all subjects and movement patterns used to perform these manual ADL. Nevertheless, most activities performed in this study were conducted in a similar manner across subjects and repetitions, which is encouraging for the future automated applications of the algorithm.

Classifying activities into different categories is an important first step, because manual activities that belong to the same category are likely to be equally impaired in a given pathology since they involve the same movement pattern. Indeed, the perceived difficulty of the activities of ABILHAND has shown that, for instance, for stroke patients, manual

activities are more challenging if they require both hands and even more challenging if they involve the fingers of both hands (Penta et al., 2001). In rheumatoid arthritis, challenging activities are those that involve higher stress at the joints, whether uni- or bi-manual (Durez et al., 2007). Therefore, for clinical follow-up of manual activity, we can hypothesize that the achievement of a type of activity is likely a very good indicator of recovery. In addition, some activities are usually only seldom performed during the day (e.g., "tying shoelaces" and "buttoning a shirt"), and grouping them as categories allows continuous monitoring whatever actual activities are performed during the day. Another argument for grouping the activities is the ability to target patients with different occupational profiles. For example, an office worker would spend most of the day typing on a keyboard or writing, while a manual worker would, rather, manipulate tools.

One strength of our study is the selection of activities that have been shown to characterize manual ability in patients with various pathologies (Durez et al., 2007; El Khoury et al., 2020b; Penta et al., 2001). The possibility to recognize the activity categories, or, in a later step, the execution of these individual activities in daily life will pave the way for comparisons between the patient-reported questionnaire scores and objective automated monitoring. Indeed, the correlations observed between the kinematic analysis of the upper limb, questionnaire scores and observational methods (Patel et al., 2010; Subramanian et al., 2010) indicate that an approach combining objective activity monitoring and questionnaire scores could help clinicians in the selection of the optimal treatments for their patients. Using such a combined approach, clinicians will better discern between capability, which describes what the patient can do in their daily environment, from performance, which refers to what the patient actually does (Holsbeeke et al., 2009).

The upper extremity is conceptualized as a single functional unit with the shoulder, elbow and wrist joints used to position the end-effector organ,

the hand, in space. The chosen localizations for the sensors allowed the capturing of the functioning of both hands very well. The wrist sensors are able to measure the movement of the hand in space, while the finger sensors record the movement of the fingers. The presence of sensors on the thumb and index finger allowed our device to be sensitive to movements of the hand involving different types of pinches and grasps (e.g., writing and handling tools), as well as activities involving fine finger movements (e.g., typing) (Napier, 1956). The addition of sensors on other locations, such as the third finger and the fingernails for precise manipulation, and the fifth finger for power grasping, could possibly provide more information regarding the type of movement. However, this additional information would come at the cost of obtrusiveness and a plethora of data. The number of sensors used in the current study is higher than in similar studies dealing with recognizing activities of the upper limb (Biswas et al., 2014; Lemmens et al., 2015). However, they provide a very good amount of data for the development of a more complex algorithm, and their location corresponds to that of everyday accessories (rings and wristbands), allowing the definitive monitoring device to be unobtrusive and ergonomic.

Cut-off values were found to be quite similar across the different analyses, except for that of the fourth participant, whose exclusion yielded slightly different results. The stability of the HR and FWR is promising regarding the generalization of the algorithm to a larger population. Participants performed the ADL as they would do in their normal life and with objects of their home environment. Unconstraining the experiment in this manner helped to generate a wide range of variability in the data, which could ultimately result in the development of an algorithm that is more readily applicable in real life. We obtained very good results in spite of potential measurement errors due to the small displacement of the sensor over the skin.

Commonly used pattern recognition approaches are neural networks, structural matching, template matching and statistical classification (Jain et

al., 2000; Preece et al., 2009). The latter approach was used in the present paper, in which each pattern is represented in terms of features of measurement. This has proven effective in developing a simple algorithm for hand activities' classification. Results are encouraging and show that activities can be reliably detected in normal subjects performing unconstrained movements. Future research should include a larger sample size to test for the stability of our chosen cutoffs, and testing of the algorithm on patients with an impaired hand function. Improvements in the algorithm could be made by using artificial intelligence (e.g., machine learning and pattern recognition), which could ultimately distinguish between individual activities. With these improvements, one should be able to determine the benefit of recognizing individual activities compared to categories. The simpler process of categorizing activities might prove sufficient for clinical applications. Nonetheless, substantial impairments can alter the execution of an activity through compensatory mechanisms, and more sophisticated algorithms might prove more appropriate in this case. A critical development of the current prototype would be an extension to a wireless system connected to a smartphone. This would allow recognition of manual activities as well as the context in which they are carried out (e.g., while sitting or walking). Recognizing the beginning and end of an activity was not addressed in this paper, but will be an essential step for the future implementation of the monitoring device in real life.

Using the monitoring device in combination with the questionnaires, the clinician will be able to optimize the patient's treatment and follow-up. A clinical improvement should manifest into more hand use and, thus, more ADL recorded on the device, as well as higher scores on the questionnaires due to a decrease in perceived difficulty. The physician will also be able to personalize the patient's therapy by tracking and focusing on a particular activity that is judged as important for the patient.

Ultimately, we aim at developing a manual activity monitoring device with wireless sensors and an autonomous power supply in order to capture

manual activities in the patient's natural environment. The compatibility of the chosen locations for the sensors with everyday life accessories will not hinder the execution of ADL. Gathering objective data from this device could be combined with patient-reported data from questionnaires in order to provide a comprehensive and global approach for outcome evaluation, clinical decision-making, patient monitoring and the tracking of rehabilitation progress.

CHAPTER 6

Discussion and future research

6.1. Summary of contributions

In this thesis, we first validated the ABILHAND questionnaire for use in hand surgery, using the Rasch model. A preliminary 90-item questionnaire was presented to 216 patients representing the diagnoses most frequently encountered in hand surgery, including distal radius fracture, basal thumb arthritis, carpal tunnel syndrome, and heavy wrist surgery. The Rasch model was used to select 23 items (mostly bimanual) that constitute the ABILHAND-HS questionnaire. The obtained scale is a linear, invariant and unidimensional continuum that defines manual ability. Unidimensionality means that only one construct (i.e. manual ability) is measured by the instrument. The scale is linear, which means that the scale unit, the logit, is continuous and presents a fixed increment along the whole scale. This allows comparing and quantifying patient improvement. Invariance means that the measures are unbiased with respect to patients' characteristics (such as age, gender or involved hand) other than the one being assessed, i.e. manual ability.

We then showed through a clinical case how the scale can be used in clinical practice for patients' evaluation. Patient's responses can be entered into a form available on the website www.rehab-scales.org to convert the questionnaire raw scores into manual ability measures. The web routine also gives a visual representation of the patient's answers, as well as patient location in logits and centiles, confidence interval and outlier answers. The items have an established difficulty hierarchy, which can provide clinicians with useful information regarding the activities that patients can or cannot do, so that treatment goals are appropriately challenging for the patient.

One important property of a scale is its ability to detect change. The responsiveness of the ABILHAND-HS has been studied, and its minimal clinically important difference has been defined for use in clinical studies and practice. Eighty-seven patients were assessed multiple times with the ABILHAND-HS questionnaire and with a global rating of change scale

(GRCS). Responsiveness was tested according to both group-level and individual-level approaches. Mean score change, effect size, standardized response mean and reliable change index were calculated for groups of patients according to their GRCS. The responsiveness indices showed that the change in manual ability measures was higher in patients who reported a great improvement in their perceived status. On an individual level, the proportion of patients with a significant improvement was higher with the increase in the GRCS. A minimal clinically important difference of 0.50 logits (4.2 centiles) was determined based on the ROC curve, the value corresponding to a small effect size and the value of the change score in the "minimal improvement" group. The ABILHAND-HS questionnaire showed a good sensitivity to change and can thus be used for the evaluation of the effect of treatments for hand surgery and in a research setting.

Finally, we showed that a simple algorithm is capable of classifying manual activities of daily living into categories using data from wrist- and finger-worn inertial sensors. Six participants without pathology of the upper limb performed 14 activities of daily living. Gyroscope signals were used to analyze the angular velocity pattern for each activity. The elaboration of the algorithm was based on the examination of the activity at the different levels (hand, fingers and wrist) and the relationship between them for the duration of the activity. A leave-one-out cross-validation was used to validate our algorithm. The algorithm allowed the classification of manual activities into five different categories through three consecutive steps, based on hands ratio (i.e. activity of one or both hands) and fingers-to-wrist ratio (i.e. finger movement independently of the wrist). On average, the algorithm made the correct classification in 87.4% of cases. The proposed algorithm has high overall accuracy, yet its computational complexity is very low as it involves only averages and ratios, making its interpretation very intuitive, which is important to guarantee its potential use in clinical practice.

6.2. Future directions

The ABILHAND-HS scale described in chapter 2 has been validated in a sample of 216 patients with four diagnostic groups. Future research should aim at verifying item calibrations and their stability in a larger sample of hand surgery patients. The scale should also be validated in more diagnostic groups such as ligament injuries and tendinopathies. Hospitals are switching to electronic medical records to improve the quality of health care. One major advantage of this transformation would be the integration of electronic questionnaires into patients' medical records (Franklin et al., 2017). This would allow data coming from a big number of patients to be effortlessly collected. Such large-scale calibrations would refine our initial findings and confirm the applicability of the ABILHAND-HS to most hand surgery patients.

We built a prototype device (InSense©, Arsalis, Belgium) to capture human activity signals using inertial measurement units. Using the angular velocity readings from the gyroscopes, we were able to develop an algorithm that classifies manual activities into five categories (see Chapter 5). Many steps are still needed to achieve a fully operational hand activities monitoring device. Improvements of the algorithm could be made to allow recognizing individual activities. This step could be achieved by using artificial intelligence (e.g. machine learning, pattern recognition). The main input would be a large database of signals recorded during identified relevant activities. One would select the most relevant signals that allow accurate discrimination among these activities. One challenge would be to find a trade-off between classification accuracy and the number of signals used. The goal is to find the lowest number of signals – hence a potentially minimal device footprint – that allows a reasonable classification success rate. The hand activity monitoring device will embed artificial intelligence to (1) identify the activity performed by the user and (2) extract indices that can quantify how well each activity is performed.

The device will be made more ergonomic with wireless sensors and an autonomous power supply in order to capture manual activities in the patient's natural environment. The incorporation of sensors into everyday life accessories (such as rings and wristbands or watches) will allow testing the device in the patient's natural environment, as they will not hinder the execution of activities of daily living.

In a daily life setting, the data from the new prototype device will first be collected on a limited number of patients and at a single time point to confirm the device operability in realistic situations of daily life. In contrast with the laboratory setting in which patients can adapt their behavior and performance due to the specificity of the context, the daily life setting will be much more ecological and realistic, albeit with more frequent compensation strategies at home than in the lab. At this stage, it will be very useful to compare the performance measured by the new prototype device with the patient-reported ABILHAND scores and with the gold standard expert observer. The ABILHAND questionnaire measures manual ability as a latent variable, while the monitoring device measures it as an observable variable. Comparing both results will inform us whether the measures taken with the two tools are two facets of the same construct or not. If the two tools measure the same construct, the most cost-effective one would be adopted for routine patient evaluation. If they show some divergence, comparing them with the gold standard expert observer would provide some insights about the different aspects they measure, and they could thus be used in combination for a more comprehensive clinical evaluation. This task will also consist in collecting patients' feedback on the device ergonomics and operability in view of the ergonomic validation of the prototype device.

6.3. Contribution of outcome measures to patient care

As discussed throughout this thesis, patient-reported outcome measures are valuable and crucial tools to assess patients' functioning and

to evaluate different treatment options. They have an already well established place in research as the primary outcome measure in most clinical studies. However, their implementation in clinical practice and adoption by physicians remains scarce, and these are not consistently used in decision making (MacDermid, 2014). The reasons for this might be the limited time in consultation to read and interpret the results, the lack of familiarity of the physicians with these instruments, and the lack of knowledge to interpret the results.

At the patient level, data from PROMs can help clinicians guide their patients by providing them with crucial information. They allow patients to understand what to expect during recovery (Baumhauer, 2017). For example, patients who undergo surgery often want to know when they can return to work or resume their sports activities. By collecting prospective data gathered at the population level, we could create a roadmap of recovery that describes the natural evolution as a function of the pre-operative score. We could compare an individual patient's preoperative score with this global data to give more accurate personalized predictions. This could help answer patients' questions and set appropriate expectations. PROM data can also be used to minimize variation in patient care. For example, institutions could compare data from different surgical procedures performed for the same condition, in order to determine which one(s) have the best outcomes from the patient's perspective. For procedures with similar outcomes, other factors such as costs, risks, and time to full recovery after surgery could be compared. When certain procedures are found to have less favorable outcomes, institutions could determine whether an individual surgeon's technique needs improvement or if the treatment approach should be abandoned completely (Baumhauer, 2017). With such a strategy, surgeons could identify areas where they need improvement, eliminate procedures with less favorable outcomes, and avoid performing surgeries on patients who are unlikely to benefit from them. It could also enhance patient satisfaction with care by helping physicians set appropriate expectations

regarding a patient's return to work, school, or sports. Most importantly, PROMs place the patient's voice at the forefront of health care delivery.

As stated in the introduction, the trend in the reimbursement of treatments by private or social insurances is shifting from "pay for an act" to "pay for results" (Porter, 2009). Therefore, current healthcare system reforms are focusing on increasing patient value, which is defined as the health outcomes achieved per unit of currency (Porter, 2010)

$$Patient\ value = \frac{Health\ outcomes}{Cost}$$

This goal is what matters most to patients and unites the interests of all actors in the healthcare system. The ABILHAND-HS could be used as an outcome in hand surgery to further implement the concept of value. Indeed, the logit that is used to measure manual ability is linear, which means that the distance between units is constant throughout the measurement range. This could open the way for the development of a new index based on the change in manual ability in logits, divided by the cost:

$$Value\ in\ hand\ surgery = \frac{Improvement\ (in\ logits)}{Cost}$$

For a given pathology that has several treatment options, the treatment that gives the best value (the best increase in manual ability while minimizing the costs) should be favored by hand surgeons and reimbursed third-party payers. The concept of value, as presented with manual ability as an outcome, can also be applied to other important outcomes in hand surgery such as grip strength, return to work, complication rates or pain reduction.

Valuable data could be gathered using a hand activities monitoring device, for all conditions that affect the upper extremity. This could be in the form of quantitative data (i.e. the amount of hand use) or qualitative data: which activities the patient does or does not perform. Analysis of such data could provide insight into the type of impairment that a patient has. Loss of

strength for example can manifest through the impossibility of performing tasks such as opening a jar. The absence of activities requiring grip strength can prompt the clinician to question the patient about the reason such activities have not been performed. A surgical procedure destined to improve the patient's hand function should have a noticeable impact on the patient's daily life activities, and this could be tracked with the monitoring device in the postoperative period. Such data could be gathered on a large scale to determine which procedure(s) have the most impact on the restoration of daily life activities.

The objective aspect of the device could provide insight into cases where the clinical examination and diagnostic tests do not align with the patients' report of symptoms and limitations in their daily life activities. By confirming the patient's claims, it could reinforce the physician-patient relationship.

6.4. Conclusion

The implementation of patient-centered outcomes throughout the medical field in general, and throughout hand surgery in particular, is advantageous for patients and clinicians alike. These measures are critical components of numerous aspects of health care delivery, including shared decision-making, post-intervention monitoring, research, quality, and value-based health care (Makhni et al., 2021). Patient-reported outcome measures and a manual activities monitoring device, as developed in this thesis, have the potential to improve medical practice by promoting a patient-centered model of healthcare delivery. We hope that this work lays the foundation for many research projects in the field of outcome measurement.

Bibliography

Aktekin LA, Eser F, Başkan BM, et al. Disability of Arm Shoulder and Hand Questionnaire in rheumatoid arthritis patients: relationship with disease activity, HAQ, SF-36. Rheumatol Int. 2011, 31: 823–6.

Alderman AK, Chung KC. Measuring Outcomes in Hand Surgery. Clinics in Plastic Surgery. 2008, 35: 239–50.

Amorosa LF, Vitale MA, Brown S, Kaufmann RA. A Functional Outcomes Survey of Elderly Patients who Sustained Distal Radius Fractures. Hand (New York, N,Y). 2011, 6: 260–7.

Andrich D. Category ordering and their utility. Rasch Measurement Transactions. 1996, 9: 464–5.

Andrich D. *Rasch Models for Measurement*. 2455 Teller Road, Thousand Oaks California 91320 United States of America, SAGE Publications, Inc., 1988.

Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978, 43: 561–73.

Andrich D, Sheridan B, Luo G. Polytomous data. *Interpreting RUMM2030: part II.* 5th ed. Western Australia, Perth, 2013.

Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014, 34: 502–8.

Arnould C, Penta M, Renders A, Thonnard J-L. ABILHAND-Kids: a measure of manual ability in children with cerebral palsy. Neurology. 2004, 63: 1045–52.

Arnould C, Penta M, Thonnard J. Hand impairments and their relationship with manual ability in children with cerebral palsy. Journal of Rehabilitation Medicine. 2007, 39: 708–14.

Arnould C, Vandervelde L, Batcho CS, Penta M, Thonnard J-L. Can manual ability be measured with a generic ABILHAND scale? A cross-sectional study conducted on six diagnostic groups. BMJ Open. 2012, 2: e001807.

Barbier O, Penta M, Thonnard J-L. Outcome evaluation of the hand and wrist according to the International Classification of Functioning, Disability, and Health. Hand Clinics. 2003, 19: 371–8.

Batcho CS, Durez P, Thonnard J-L. Responsiveness of the ABILHAND questionnaire in measuring changes in rheumatoid arthritis patients. Arthritis Care & Research. 2011, 63: 135–41.

Baumhauer JF. Patient-Reported Outcomes—Are They Living Up to Their Potential? N Engl J Med. 2017, 377: 6–9.

Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol. 2001, 54: 1204–17.

Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: Reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. Journal of Clinical Epidemiology. 1997, 50: 79–93.

Beaton DE, Wright JG, Katz JN. Development of the QuickDASH: Comparison of Three Item-Reduction Approaches. Journal of Bone and Joint Surgery (American). 2005, 87A: 9.

Berkanovic E, Hurwicz M-L, Lachenbruch PA. Concordant and discrepant views of patients' physical functioning. Arthritis & Rheumatism. 1995, 8: 94–101.

Biswas D, Corda D, Baldus G, et al. Recognition of elementary arm movements using orientation of a tri-axial accelerometer located near the wrist. Physiological Measurement. 2014, 35: 1751–68.

Bobos P, Lalone EA, Grewal R, MacDermid JC. Do Impairments Predict Hand Dexterity After Distal Radius Fractures? A 6-Month Prospective Cohort Study. HAND. 2018, 13: 441–7.

Brentani E, Golia S. Unidimensionality in the Rasch model: how to detect and interpret. Statistica. 2007, 67: 253–61.

Cano SJ, O'Connor RJ, Thompson AJ, Hobart JC. Exploring disability rating scale responsiveness II: do more response options help? Neurology. 2006, 67: 2056–9.

Cella D, Bullinger M, Scott C, Barofsky I, Clinical Significance Consensus Meeting Group. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. Mayo Clin Proc. 2002, 77: 384–92.

Chang W-C, Chan C. Rasch analysis for outcomes measures: some methodological considerations. Archives of Physical Medicine and Rehabilitation. 1995, 76: 934–9.

Changulani M, Okonkwo U, Keswani T, Kalairajah Y. Outcome evaluation measures for wrist and hand – which one to choose? International Orthopaedics. 2008, 32: 1–6.

Chatterjee JS, Price PE. Comparative responsiveness of the Michigan Hand Outcomes Questionnaire and the Carpal Tunnel Questionnaire after carpal tunnel release. J Hand Surg Am. 2009, 34: 273–80.

Chen CC, Granger CV, Peimer CA, Moy OJ, Wald S. Manual Ability Measure (MAM-16): a preliminary report on a new patient-centred and task-oriented outcome measure of hand function. Journal of Hand Surgery. 2005, 30: 207–16.

Chen S-R, Shen Y-P, Ho T-Y, et al. One-Year Efficacy of Platelet-Rich Plasma for Moderate-to-Severe Carpal Tunnel Syndrome: A Prospective, Randomized, Double-Blind, Controlled Trial. Arch Phys Med Rehabil. 2021, 102: 951–8.

Chen W-H, Lenderking W, Jin Y, Wyrwich KW, Gelhorn H, Revicki DA. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. Qual Life Res. 2014, 23: 485–93.

Christensen L, Mendoza JL. A method of assessing change in a single subject: An alteration of the RC index. Behavior Therapy. 1986, 17: 305–8.

Chung KC, Pillsbury MS, Walters MR, Hayward RA. Reliability and validity testing of the Michigan Hand Outcomes Questionnaire. The Journal of Hand Surgery. 1998, 23: 575–87.

Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. Arch Phys Med Rehabil. 2008, 89: 69–74.

Coenen M, Kus S, Rudolf K-D, et al. Do patient-reported outcome measures capture functioning aspects and environmental factors important to individuals with injuries or disorders of the hand? Journal of Hand Therapy. 2013, 26: 332–42.

Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J, L. Erlbaum Associates, 1988.

Conrad KJ, Smith EV. International conference on objective measurement: applications of Rasch analysis in health care. Med Care. 2004, 42: I1-6.

Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J. 2007, 7: 541–6.

Cowie J, Anakwe R, McQueen M. Factors Associated with One-Year Outcome after Distal Radial Fracture Treatment. Journal of Orthopaedic Surgery. 2015, 23: 5.

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 2003, 56: 395–407.

Cuesta-Vargas AI, Galán-Mercant A, Williams JM. The use of inertial sensors system for human motion analysis. Phys Ther Rev. 2010, 15: 462–73.

Desrosiers J, Noreau L, Rochette A, Bourbonnais D, Bravo G, Bourget A. Predictors of long-term participation after stroke. Disabil Rehabil. 2006, 28: 221–30.

DeVellis RF. Classical test theory. Med Care. 2006, 44: S50-59.

Dubert T. Outcome measurements in hand and upper limb surgery. Chirurgie de La Main. 2014, 33: 235–46.

Durez P, Fraselle V, Houssiau F, Thonnard J, Nielens H, Penta M. Validation of the ABILHAND questionnaire as a measure of manual ability in patients with rheumatoid arthritis. Ann Rheum Dis. 2007, 66: 1098–105.

El Khoury G, Barbier O, Libouton X, Thonnard J-L, Lefevre P, Penta M. Manual ability in hand surgery patients: validation of the ABILHAND scale in four diagnostic groups. MedRxiv. 2020a: 2020.07.02.20144147.

El Khoury G, Barbier O, Libouton X, Thonnard J-L, Lefèvre P, Penta M. Manual ability in hand surgery patients: Validation of the ABILHAND scale in four diagnostic groups. PLoS One. 2020b, 15: e0242625.

Esakki S, MacDermid JC, Vincent JI, Packham TL, Walton D, Grewal R. Rasch analysis of the patient-rated wrist evaluation questionnaire. Arch Physiother. 2018, 8.

Fisher W. Reliability, Separation, Strata Statistics. Rasch Meas Trans. 1992, 6: 238.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess. 1998, 2: i–iv, 1–74.

Franchignoni F, Ferriero G, Giordano A, Sartorio F, Vercelli S, Brigatti E. Psychometric properties of QuickDASH – A classical test theory and Rasch analysis study. Manual Therapy. 2011, 16: 177–82.

Franchignoni F, Giordano A, Sartorio F, Vercelli S, Pascariello B, Ferriero G. Suggestions for Refinement of the Disabilities of the Arm, Shoulder and Hand Outcome Measure (DASH): A Factor Analysis and Rasch Validation Study. Archives of Physical Medicine and Rehabilitation. 2010, 91: 1370–7.

Franchignoni F, Vercelli S, Giordano A, Sartorio F, Bravini E, Ferriero G. Minimal Clinically Important Difference of the Disabilities of the Arm, Shoulder and Hand Outcome Measure (DASH) and Its Shortened Version (QuickDASH). J Orthop Sports Phys Ther. 2014, 44: 30–9.

Franklin P, Chenok K, Lavalee D, et al. Framework To Guide The Collection And Use Of Patient-Reported Outcome Measures In The Learning Healthcare System. EGEMS (Wash DC). 2017, 5: 17.

Giladi AM, Chung KC. Measuring Outcomes in Hand Surgery. Clinics in Plastic Surgery. 2013, 40: 313–22.

Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice? Journal of Rehabilitation Medicine. 2012, 44: 97–8.

Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol. 1989, 42: 403–8.

Hagell P, Westergren A. Sample Size and Statistical Conclusions from Tests of Fit to the Rasch Model According to the Rasch Unidimensional Measurement Model (Rumm) Program in Health Outcome Measurement. J Appl Meas. 2016, 17: 416–31.

Hagquist C, Andrich D. Recent advances in analysis of differential item functioning in health research using the Rasch model. Health and Quality of Life Outcomes. 2017, 15: 181.

Halilaj E, Rajagopal A, Fiterau M, Hicks JL, Hastie TJ, Delp SL. Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. Journal of Biomechanics. 2018, 81: 1–11.

Hewlett SA. Patients and Clinicians Have Different Perspectives on Outcomes in Arthritis. J Rheumatol. 2003, 30: 877–9.

Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol. 2007, 6: 1094–105.

Holland PW, Wainer H. *Differential item functioning*. Hillsdale, NJ, US, Lawrence Erlbaum Associates, Inc, 1993.

Holsbeeke L, Ketelaar M, Schoemaker MM, Gorter JW. Capacity, Capability, and Performance: Different Constructs or Three of a Kind? Archives of Physical Medicine and Rehabilitation. 2009, 90: 849–55.

Hong SW, Gong HS, Park JW, Roh YH, Baek GH. Validity, Reliability and Responsiveness of the Korean Version of Quick Disabilities of the Arm,

Shoulder, and Hand Questionnaire in Patients with Carpal Tunnel Syndrome. J Korean Med Sci. 2018, 33: e249.

Hosmer DW, Lemeshow S. *Applied Logistic Regression*. John Wiley & Sons, 2004.

Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). Am J Ind Med. 1996, 29: 602–8.

ICHOM. ICHOM | About Us | Our Mission | History | Michael Porter. ICHOM. 2021a. https://www.ichom.org/mission/ (accessed July 14, 2021).

ICHOM. ICHOM | ICHOM Standard Sets | View Our Collection. ICHOM. 2021b. https://www.ichom.org/standard-sets/ (accessed September 16, 2021).

INAMI. Une rééducation via application mobile après la pose d'une prothèse du genou ou de la hanche - INAMI. 2021. https://www.riziv.fgov.be/fr/themes/cout-remboursement/Pages/remboursement-soins-distance-reeducation-prothese-genou-hanche-application-mobile.aspx (accessed July 14, 2021).

Jacobson NS, Follette WC, Revenstorf D. Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. Behavior Therapy. 1984, 15: 336–52.

Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991, 59: 12–9.

Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Controlled Clinical Trials. 1989, 10: 407–15.

Jain AK, Duin RPW, Jianchang Mao. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000, 22: 4–37.

Jaremko JL, Lambert RGW, Rowe BH, Johnson JA, Majumdar SR. Do radiographic indices of distal radius fracture reduction predict outcomes in older adults receiving conservative treatment? Clinical Radiology. 2007, 62: 65–72.

Jerosch-Herold C, Leite JC de C, Song F. A systematic review of outcomes assessed in randomized controlled trials of surgical interventions for carpal tunnel syndrome using the International Classification of Functioning, Disability and Health (ICF) as a reference tool. BMC Musculoskeletal Disorders. 2006, 7.

Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. J Clin Epidemiol. 1994, 47: 81–7.

Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. Journal of Manual & Manipulative Therapy. 2009, 17: 163–70.

Kazis LE, Anderson JJ, Meenan RF. Effect Sizes for Interpreting Changes in Health Status: Medical Care. 1989, 27: S178–89.

Klum M, Wolf MB, Hahn P, Leclère FM, Bruckner T, Unglaub F. Normative Data on Wrist Function. The Journal of Hand Surgery. 2012, 37: 2050–60.

Körver RJP, Senden R, Heyligers IC, Grimm B. Objective outcome evaluation using inertial sensors in subacromial impingement syndrome: a five-year follow-up study. Physiol Meas. 2014, 35: 677–86.

Kotsis SV, Chung KC. Responsiveness of the Michigan Hand Outcomes Questionnaire and the Disabilities of the Arm, Shoulder and Hand questionnaire in carpal tunnel surgery. J Hand Surg Am. 2005, 30: 81–6.

Kwakkel G, Kollen BJ, van der Grond J, Prevo AJH. Probability of regaining dexterity in the flaccid upper limb: impact of severity of paresis and time since onset in acute stroke. Stroke. 2003, 34: 2181–6.

Lemmens RJM, Janssen-Potten YJM, Timmermans AAA, Smeets RJEM, Seelen HAM. Recognizing Complex Upper Extremity Activities Using Body Worn Sensors. PLOS ONE. 2015, 10: e0118642.

Levine DW, Simmons BP, Koris MJ, et al. A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. JBJS. 1993, 75: 1585–92.

Liang MH, Fossel AH, Larson MG. Comparisons of Five Health Status Instruments for Orthopedic Evaluation. Medical Care. 1990, 28: 632–42.

Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas. 2002, 3: 85–106.

Linacre JM. Detecting multidimensionality: which residual data-type works best? J Outcome Meas. 1998, 2: 266–83.

Lord F, Novick M. Statistical Theories of Mental Test Scoires. Addison-Westley Publ. Co. Reading, Mass. 1968.

MacDermid JC. Patient-Reported Outcomes. Hand Clinics. 2014, 30: 293–304.

MacDermid JC, Tottenham V. Responsiveness of the disability of the arm, shoulder, and hand (DASH) and patient-rated wrist/hand evaluation (PRWHE) in evaluating change after hand therapy. J Hand Ther. 2004, 17: 18–23.

MacDermid JC, Turgeon T, Richards RS, Beadle M, Roth JH. Patient rating of wrist pain and disability: a reliable and valid measurement tool. J Orthop Trauma. 1998, 12: 577–86.

Makhni EC, Swantek AJ, Ziedas AC, et al. The Benefits of Capturing PROMs in the EMR. NEJM Catalyst. 2021, 2.

Masters GN. A rasch model for partial credit scoring. Psychometrika. 1982, 47: 149–74.

McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Quality of Life Research. 1995, 4: 293–307.

Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Archives of Physical Medicine and Rehabilitation. 1989, 70: 308–12.

Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Medicine. 1978, 8: 283–98.

Napier JR. The prehensile movements of the human hand. The Journal of Bone and Joint Surgery British Volume. 1956, 38-B: 902–13.

van Nes SI, Vanhoutte EK, van Doorn PA, et al. Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. Neurology. 2011, 76: 337–45.

Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997, 50: 869–79.

Owen DH, Agius PA, Nair A, Perriman DM, Smith PN, Roberts CJ. Factors predictive of patient outcome following total wrist arthrodesis. The Bone & Joint Journal. 2016, 98-B: 647–53.

Packham T, MacDermid JC. Measurement properties of the Patient-Rated Wrist and Hand Evaluation: Rasch analysis of responses from a traumatic hand injury population. Journal of Hand Therapy. 2013, 26: 216–24.

Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol. 2007, 46: 1–18.

Patel S, Hughes R, Hester T, et al. A Novel Approach to Monitor Rehabilitation Outcomes in Stroke Survivors Using Wearable Technology. IEEE. 2010.

Penta M, Tesio L, Arnould C, Zancan A, Thonnard J-L. The ABILHAND Questionnaire as a Measure of Manual Ability in Chronic Stroke Patients: Rasch-Based Validation and Relationship to Upper Limb Impairment. Stroke. 2001, 32: 1627–34.

Penta M, Thonnard J-L, Tesio L. ABILHAND: a Rasch-built measure of manual ability. Archives of Physical Medicine and Rehabilitation. 1998, 79: 1038–42.

Porter ME. What Is Value in Health Care? New England Journal of Medicine. 2010, 363: 2477–81.

Porter ME. A strategy for health care reform--toward a value-based system. N Engl J Med. 2009, 361: 109–12.

Preece SJ, Goulermas JY, Kenney LPJ, Howard D, Meijer K, Crompton R. Activity identification using body-mounted sensors—a review of classification techniques. Physiological Measurement. 2009, 30: R1–33.

de Putter CE, Selles RW, Polinder S, Panneman MJM, Hovius SER, van Beeck EF. Economic impact of hand and wrist injuries: health-care costs and

productivity costs in a population-based study. J Bone Joint Surg Am. 2012, 94: e56.

Ramp M, Khan F, Misajon RA, Pallant JF. Rasch analysis of the Multiple Sclerosis Impact Scale (MSIS-29). Health Qual Life Outcomes. 2009, 7: 58.

Rasch G. *Probabilistic models for some intelligence and attainment tests*. Chicago, IL, The University of Chicago Press, 1980.

Rau G, Disselhorst-Klug C, Schmidt R. Movement biomechanics goes upwards: from the leg to the arm. Journal of Biomechanics. 2000, 33: 1207–16.

Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology. 2008, 61: 102–9.

Rowe JB, Friedman N, Bachman M, Reinkensmeyer DJ. The Manumeter: a non-obtrusive wearable device for monitoring spontaneous use of the wrist and fingers. IEEE Int Conf Rehabil Robot. 2013, 2013: 6650397.

Rowe JB, Friedman N, Chan V, Cramer SC, Bachman M, Reinkensmeyer DJ. The variable relationship between arm and hand use: a rationale for using finger magnetometry to complement wrist accelerometry when measuring daily use of the upper extremity. *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014: 4087–90.

Roy SH, Cheng MS, Chang S, et al. A Combined sEMG and Accelerometer System for Monitoring Functional Activity in Stroke. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2009, 17: 585–94.

Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining Clinically Important Differences in Health Status Measures. Pharmacoeconomics. 1999, 15: 141–55.

Sana Furrukh, Isselbacher Eric M., Singh Jagmeet P., Heist E. Kevin, Pathik Bhupesh, Armoundas Antonis A. Wearable Devices for Ambulatory Cardiac Monitoring. Journal of the American College of Cardiology. 2020, 75: 1582–92.

Schwarz N, Sudman S (Eds.). *Autobiographical Memory and the Validity of Retrospective Reports*. New York, Springer-Verlag, 1994.

da Silva NC, Chaves TC, Dos Santos JB, et al. Reliability, validity and responsiveness of Brazilian version of QuickDASH. Musculoskelet Sci Pract. 2020, 48: 102163.

Sloan J, Symonds T, Vargas-Chanes D, Fridley B. Practical Guidelines for Assessing the Clinical Significance of Health-Related Quality of Life Changes within Clinical Trials. Ther Innov Regul Sci. 2003, 37: 23–31.

Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas. 2002, 3: 205–31.

Smith EVJ, Conrad KM, Chang K, Piazza J. An Introduction to Rasch Measurement for Scale Development and Person Assessment. J Nurs Meas. 2002, 10: 189–206.

Spearman C. "General Intelligence," Objectively Determined and Measured. The American Journal of Psychology. 1904, 15: 201–92.

Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol. 1996, 49: 711–7.

Subramanian SK, Yamanaka J, Chilingaryan G, Levin MF. Validity of movement pattern kinematics as measures of arm motor impairment poststroke. Stroke. 2010, 41: 2303–8.

Swiontkowski MF, Buckwalter JA, Keller RB, Haralson R. The outcomes movement in orthopaedic surgery: where we are and where we should go. The Journal of Bone and Joint Surgery American Volume. 1999, 81: 732–40.

Synn AJ, Makhni EC, Makhni MC, Rozental TD, Day CS. Distal Radius Fractures in Older Patients: Is Anatomic Reduction Necessary? Clin Orthop Relat Res. 2009, 467: 1612–20.

Szabo RM. Outcomes assessment in hand surgery: When are they meaningful? The Journal of Hand Surgery. 2001, 26: 993–1002.

Tamura T. Wearable Inertial Sensors and Their Applications. *Wearable Sensors*. Elsevier, 2014: 85–104.

Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis & Rheumatism. 2007, 57: 1358–62.

Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007, 60: 34–42.

Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of Rehabilitation Medicine. 2003, 35: 105–15.

Thanawattano C, Pongthornseri R, Anan C, Dumnin S, Bhidayasiri R. Temporal fluctuations of tremor signals from inertial sensor: a preliminary study in differentiating Parkinson's disease from essential tremor. Biomed Eng Online. 2015, 14: 101.

Thurstone L. *The measurement of values*. Chicago, The University of Chicago Press, 1959.

Turner D, Schünemann HJ, Griffith LE, et al. The minimal detectable change cannot reliably replace the minimal important difference. Journal of Clinical Epidemiology. 2010, 63: 28–36.

Turner D, Schünemann HJ, Griffith LE, et al. Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. Journal of Clinical Epidemiology. 2009, 62: 374–9.

Vandervelde L, Van den Bergh PYK, Penta M, Thonnard JL. Validation of the ABILHAND questionnaire to measure manual ability in children and adults with neuromuscular disorders. Journal of Neurology, Neurosurgery & Psychiatry. 2010, 81: 506–12.

Vandervelde L, Van den Bergh PYK, Renders A, Goemans N, Thonnard J-L. Relationships between motor impairments and activity limitations in patients with neuromuscular disorders. Journal of Neurology, Neurosurgery & Psychiatry. 2009, 80: 326–32.

Vanthuyne M, Smith V, Arat S, et al. Validation of a manual ability questionnaire in patients with systemic sclerosis. Arthritis & Rheumatism. 2009, 61: 695–703.

Wang T, Lin K, Wu C, Chung C, Pei Y, Teng Y. Validity, responsiveness, and clinically important difference of the ABILHAND questionnaire in patients with stroke. Arch Phys Med Rehabil. 2011, 92: 1086–91.

Wang Y-C, Hart DL, Stratford PW, Mioduski JE. Baseline dependency of minimal clinically important improvement. Phys Ther. 2011, 91: 675–88.

Ward MM, Marx AS, Barry NN. Identification of clinically important changes in health status using receiver operating characteristic curves. J Clin Epidemiol. 2000, 53: 279–84.

Ware J, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996, 34: 220–33.

Ware JE, New England Medical Center Hospital., Health Institute. *SF-36 physical and mental health summary scales : a user's manual*. Boston, Health Institute, New England Medical Center, 1994.

World Health Organization (Ed.). *International classification of functioning, disability and health: ICF*. Geneva, World Health Organization, 2001.

Wright AA, Cook CE, Baxter GD, Dockerty JD, Abbott JH. A comparison of 3 methodological approaches to defining major clinically important improvement of 4 performance measures in patients with hip osteoarthritis. J Orthop Sports Phys Ther. 2011, 41: 319–27.

Wright B, Stone M. Best test design. Measurement and Statistics. 1979.

Wright B.D. Local dependency, correlations and principal components. Rasch Meas Trans 1996;10:509–11. 1996.

Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil. 1989, 70: 857–60.

Wright BD, Masters GN. *Rating scale analysis*. Chicago, Mesa Press, 1982.

Wyrwich K.W., Nienaber NA, Tierney WM, Wolinsky FD. Linking Clinical Relevance and Statistical Significance in Evaluating Intra-Individual Changes in Health-Related Quality of Life. Medical Care. 1999, 37: 469–78.

Wyrwich Kathleen W., Tierney WM, Wolinsky FD. Further Evidence Supporting an SEM-Based Criterion for Identifying Meaningful Intra-Individual Changes in Health-Related Quality of Life. Journal of Clinical Epidemiology. 1999, 52: 861–73.

Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. J Eval Clin Pract. 2000, 6: 39–49.

Yang C-C, Hsu Y-L. A review of accelerometry-based wearable motion detectors for physical activity monitoring. Sensors (Basel). 2010, 10: 7772–88.

Young BT, Rayan GM. Outcome following nonoperative treatment of displaced distal radius fractures in low-demand patients older than 60 years. The Journal of Hand Surgery. 2000, 25: 19–28.

Zhou H, Stone T, Hu H, Harris N. Use of multiple wearable inertial sensors in upper limb motion tracking. Med Eng Phys. 2008, 30: 123–33.

Ziebland S, Fitzpatrick R, Jenkinson C. Tacit models of disability underlying health status instruments. Social Science & Medicine. 1993, 37: 69–75.