Title: A study to assess Content Validity and Inter- and Intra-rater Reliability of the Inclusion Body Myositis Functional Rating Scale (IBMFRS)

Running Head: CONTENT VALIDITY AND RELIABILITY OF THE IBMFRS

Abstract

Background and Objectives
Sporadic Inclusion Body Myositis (IBM) is a rare, muscle-wasting disease that negatively impacts health-related quality of life. Although a measure that has been developed to assess the impact of IBM, the IBM Functional Rating Scale (IBMFRS) has limited evidence of content validity or reliability, and what constitutes a meaningful change threshold; this study was conducted to address these gaps.

Methods
Adult patients with a clinical diagnosis of IBM from the UK and disease area expert healthcare professionals (HCPs) from the US and UK took part in the study. The study consisted of five stages including phone interviews (physicians), face-to-face interviews (patients), face-to-face ratings, phone ratings, and ratings of videos using the IBMFRS.

Results
The IBMFRS adequately captures all core functional impacts of IBM, which was corroborated by both patient participants and physicians when debriefing the measure. Physicians and patient participants all thought any change on the measure would be meaningful change for a patient, either improvement or worsening. The quantitative analysis demonstrated good inter-rater reliability for face-to-face ratings (intraclass correlation coefficient (ICC) > 0.7), and for video ratings (ICC > 0.9). Intra-rater reliability was excellent for face-to-face and video ratings (ICC > 0.9). Equivalence between the modes of administration, face-to-face versus phone, was also excellent (ICC > 0.9).

Discussion
The IBMFRS is content valid in assessing the key functional impacts of IBM and any change would be meaningful. It is reliable both within and across raters, and there is equivalence between different modes of administration (face-to-face vs phone).

Key Message:

Although the IBM Functional Rating Scale (IBMFRS) has been developed to assess the impact of IBM, there is limited evidence of content validity or reliability, and what constitutes a meaningful change threshold in IBM; this study was conducted to address these gaps. Overall, it is concluded that the IBMFRS is content valid in assessing the key functional impacts of IBM. It is reliable both within and across raters, and there is equivalence of scores across both face-to-face and phone modes of administration, highlighting its feasibility for use within the context of clinical trials. A 1-category change on any items was classed as meaningful by patients and physicians.

Take-home points:

1. IBM has a substantial impact on the day-to-day lives of patients and negatively impacts health-related quality of life.
2. The IBMFRS is content valid in assessing the key functional impacts of IBM.
3. The IBMFRS is reliable both within and across raters,
4. There is equivalence of scores between face-to-face and phone modes of administration.
5. A 1-category change on any item was classed as meaningful by both patients and physicians.

INTRODUCTION

Sporadic inclusion body myositis (IBM) is the most common idiopathic inflammatory myopathy (IIM) after age 50 and more commonly in men than women (Dimachkie and Barohn 2009). It is a debilitating autoimmune and degenerative condition manifesting as progressive muscle weakness, typically initially affecting the finger flexors and/or quadriceps, leading to loss of dexterity and falls. While there is a suggestion of shortened life expectancy in IBM patients, (Naddaf, Shelly et al. 2022) several causes of death have been ascribed to dysphagia or weakness of the respiratory muscle, including malnutrition, cachexia, aspiration, or respiratory failure (Machado, Brady et al. 2013).

The prevalence of IBM is unknown in the US but is conservatively estimated at 5 to 10 cases per million, while the upper prevalence estimate is 71 per million(Dimachkie and Barohn 2014) and up to 182 per million in people aged 50 or older (Shelly, Mielke et al. 2021). A 2017 meta-analysis estimated a worldwide prevalence of between 25 and 46 cases per million (Callan, Capkun et al. 2017). There is a significant unmet need for treatment options as currently there is no approved treatment for IBM. Therefore, patients typically undergo several treatments with unapproved medications without sustained effect. Despite a clear inflammatory component in IBM, no significant effect has been found with immunosuppressive therapy, such as corticosteroids, intravenous immunoglobulin (IVIG), methotrexate, or azathioprine (Lilleker, Rietveld et al. 2017, Naddaf, Barohn et al. 2018). The mainstays of supportive treatment are physical therapy, exercise, nutritional treatment, falls prevention, treatment of dysphagia, and treatment of disease-related psychological symptoms (Alexanderson 2012).

The IBMFRS is a disease-specific Clinician Reported Outcome (ClinRO) measure that assesses a patient's ability and independence in completing 10 functional activities such as swallowing, handwriting, dressing, hygiene, walking, and climbing stairs (Jackson, Barohn et al. 2008) and is administered during a patient assessment. Responses are selected using a 5-point scale, ranging from 0 (unable to perform) to 4 (normal). The measure is scored using a total score of all 10 items. The IBMFRS was developed by Jackson et al, 2008 and the Muscle Study Group, and was derived from the Amyotrophic Lateral Sclerosis (ALS) Functional Rating Scale, (Cedarbaum and Stambler 1997) a widely accepted and used measure in ALS clinical trials.

Clinical outcome assessments (COA's) positioned as primary or secondary endpoints within clinical trials require evidence for reliability and validity prior to product labelling approval (Morrow, Ramdharry et al. 2013). In rare indications, novel methods for the psychometric evaluation of COA measures must be considered since common methods may lack the statistical power to make psychometric inferences, given the small population and increased heterogeneity of the disease experience (Revicki 2018). Although there is adequate evidence of the psychometric properties of the IBMFRS, (Jackson, Barohn et al. 2008, Morrow, Ramdharry et al. 2013, Sangha, Yao et al. 2021) there is a lack of robust evidence for content validity for its use in IBM, inter-rater and intra-rater reliability,

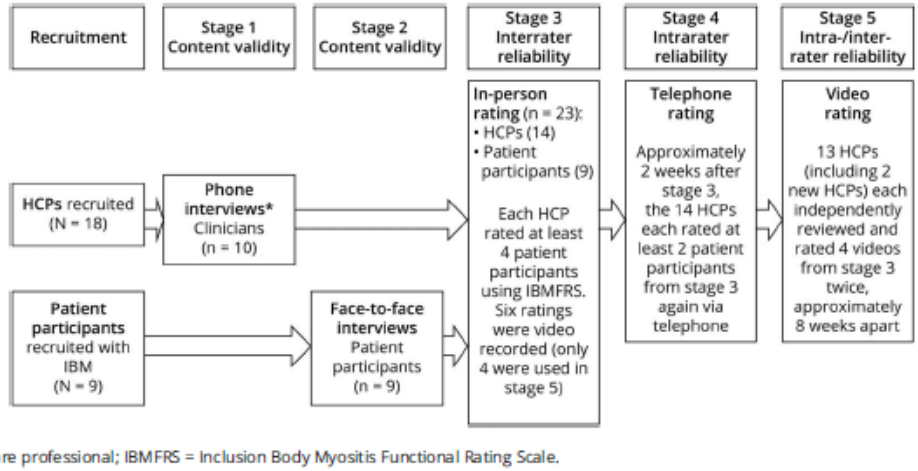
phone versus in-person administration and, for what constitutes a meaningful change. Therefore, the objective of the study was to address the identified gaps in the IBMFRS (i.e., content validity, reliability and, meaningful change threshold).

METHODS

Study design and participants

The study included adult patient participants with IBM from the UK (n = 9) and expert HCPs in IBM from the US and UK (n = 18). Expert HCPs in IBM were initially identified by two of the clinicians involved in this study (PMM and MMD). These HCPs were approached based on their clinical experience with IBM; none of the invited clinicians declined participation in this study. To be included in the panel, HCPs had to be qualified as a clinician, nurse, or physiotherapist with experience and knowledge of the history of IBM, had to have at least 2 years of experience working with IBM patients and see at least 5-10 patients a year with this condition. Inclusion criteria for patient participants required the participant to sign an informed consent form (ICF) prior to any study-related procedures. The specific inclusion criteria were: male or female, of any race aged ≥ 45 years; participant had a clinical diagnosis of IBM as determined by a specialist doctor in neuromuscular diseases and met the European Neuromuscular Centre Inclusion Body Myositis research diagnostic criteria 2011 categories for IBM (Rose 2013); could complete all study requirements and was sufficiently fluent in speaking English to participate in the interview. Participants were excluded if they had previously been or were currently taking part in an ongoing clinical trial to treat their IBM. HCPs had to be qualified as a physician, nurse, physiotherapist, or healthcare assistant with experience of the IBMFRS and IBM; have at least two years of experience in IBM and see at least 5-10 patients with IBM per year. Severity of IBM was collected via a physician-reported severity rating scale. The severity of a patient participant's IBM was defined as the perceived degree of impact of the disease on overall function i.e., the overall level of disability.

There were five stages to the study (see Fig 1). The first two stages were used to address the content validity of the IBMFRS with physicians and patients. The next three stages explored the inter and intra-rater reliability of the IBMFRS. Details of each stage are provided below.



HCP = health care professional; IBMFRS = Inclusion Body Myositis Functional Rating Scale.

Figure 1: Schematic of the study design

Stage 1 & 2: Interview study procedures (content validity)

Ten phone interviews were conducted with physicians (Stage 1) and 9 interviews were conducted with patient participants face-to-face (Stage 2). All interviews were completed by two trained qualitative

researchers using a semi-structured discussion guide (one for patient participant interviews and one for physician interviews). All interviews were recorded and transcribed verbatim. In both interviews, there were brief initial questions designed to establish a rapport and open the conversation. This was followed by a series of focused questions to explore the patient/lived experience of IBM. Patients were asked questions such as "Could you walk me through a TYPICAL day with your IBM? Start at the beginning and tell me what happens when you get up in the morning" and "Do you ever have times when your IBM is worse than usual? Can you walk me through what that is like?" The second part of the interview involved a discussion on the IBMFRS specifically. Participants provided feedback on the behaviors captured, and their importance, as well as whether the response options made sense. Alongside this, the physicians were also asked to comment on their overall understanding of the measure and ease of use. Finally, at the end of the interview, there was a discussion on meaningful change, where patients and physicians separately discussed the smallest amount of change that they would want to see on the scales for them to class this as being meaningful.

Stage 3 - 5: IBMFRS rating study procedures (inter and intra rater reliability)

All HCPs received training on how to administer the IBMFRS. For those who were interviewed, the training took place after the interview to mitigate any bias. Face-to-face ratings (stage 3) of patients occurred over a single weekend with 9 patient participants and 14 HCPs. Seven of those physicians who were interviewed also participated in Stage 3. Each of the 14 HCPs rated between 3 and 6 patient participants resulting in a total of 49 IBMFRS ratings. Six of the ratings were video recorded for use in stage 5 and involved 2 of the HCPs as raters. The videos focused only on the patient participants and were recorded by an external media company.

All 14 HCPs and the 9 patient participants who took part in Stage 3 also took part in the Stage 4 phone ratings. This involved each HCP rating at least 2 of the patient participants they had rated in Stage 3, approximately 2 weeks later, resulting in 28 ratings. Prior to the telephone rating, patient participants were asked if they had experienced a significant health event since the face-to-face rating, such as an illness or injury that might impact their rating. If a significant event was reported their data were excluded from the analysis for this stage of the study.

HCPs who took part in Stages 3 and 4 were invited to also take part in Stage 5, except the 2 HCPs whose ratings had previously been video recorded and were replaced by 2 new HCPs. One HCP decided not to participate in this stage. This resulted in a total of 13 HCPs. The six videos recorded in stage 3 were reviewed by the study team and a disease area expert to select four videos that represented a range of participants. From this, two files of the videos were developed, with each containing the same four videos, but in different orders. Therefore, at time point 1, approximately half of the sample of HCPs saw file 1, whilst the other HCPs saw file 2. At time point 2 the HCPs then viewed the alternate file. Time point 1 occurred approximately 4 weeks after stage 4 had been completed. Time point 2 then occurred approximately 8 weeks later. HCPs viewed the videos in a secure online portal and were given 14 days to review and submit their ratings.

The videos were also reviewed by an expert in IBMFRS scoring to provide a gold standard rating. This was used to compare against the HCP ratings.

The study was approved by the United Kingdom Health Research Authority (HRA) Research and the London (Surry) Research Ethics Committee (REC) (approval number 20/LO/0801).

Stages 1 & 2: Qualitative analytical methods

Transcripts of the interviews were entered into NVivo 1.0 or higher, a software package designed to facilitate the storage, coding, and analysis of qualitative data. They were coded using thematic analysis

to identify any themes, patterns, or features of interest within the data.(Braun and Clarke 2006) Saturation analysis was then undertaken on the concept elicitation data from the patient interviews only by dividing the study sample into four equal groups based on the chronological order in which they were interviewed. Saturation was considered met when no new themes or descriptions of concepts were identified in the final round of interviews.

Stages 3 - 5: Quantitative analytic methods

A power analysis to determine sample size requirements for estimating inter-rater and intra-rater reliability was computed based on a repeated measure analysis of variance (ANOVA) between and within subjects' design for the IBMFRS total score.(Qin, Nelson et al. 2019) This model assumes four observations for each rater for inter-rater reliability and two observations of the same rater over two timepoints for intra-rater reliability. Inter-rater estimates for the between-subjects parameters assume the magnitude of variability, as measured by the standardized effect size of the difference in IBMFRS scores, is ≤ 0.20 (low effect) with an inferential probability of ≥ 0.10. If observed, the variability between the physician populations at each time point will be low and non-significant, thus supporting inter-rater reliability. For estimation of the sample size intra-rater estimates for the within-subjects observations assume a correlation between repeated observations of 0.85 (high correlation). Based on these assumptions, the sample should include at least 14 raters, assessing 4 cases, over the two assessment time points to observe 80% power to detect within-group consistency and between-group variability.

Descriptive analyses of the IBMFRS items, n, frequency and percentage were computed for each of the stages where ratings were made, which was then used to support the evaluation of variability in the inter- and intra-rater agreements for the IBMFRS total score.

The total score on the IBMFRS was computed for each patient participant on each rating event (Stages 3 to 5) and for each rater. These scores were then used in fixed effect ANOVA analyses to compute intraclass correlation coefficients (ICCs) (McGraw and Wong 1996). The prespecified criterion for acceptable reliability was defined as 0.70 (range 0-1.0 with a higher threshold demonstrating better rater agreement and stability) (Landis and Koch 1977). A Many-Facets item-response analysis was conducted as a sensitivity analysis to identify a potential error in rater-reliability estimates due to rater bias (Linacre and Wright 1994). FACETS provides indices called: strata, separation, and reliability of separation (which are computed for each facet; patient, stage, timepoint, or item). The strata index indicates the number of statistically distinct levels of severity. Separation indicates the spread of the rater (or) severity levels relative to the precision of measurement. The Reliability of Separation indicates how well the data from different facets (patients, raters, stage, timepoint, items) are separated in terms of severity levels; it reflects the ability of the measure to "separate" patients (or raters) by their severity and can vary between 0 and 1.

All quantitative analyses were conducted using Statistical Analysis System (SAS) 9.4 or higher[(SAS 2013)] and FACETS 3.8.1 or higher [(Braun and Clarke 2006)].

Standard Protocol Approvals, Registrations, and Patient Consents

Written informed consent was obtained from each patient participant before enrolment into the study. All study materials were reviewed and approved by the London – Surrey Research Ethics Committee (REC reference 20/LO/0801).

Data Availability:

Anonymized data not published within this article will be made available upon reasonable request from any qualified investigator, if ethically and legally possible.

RESULTS

Patient participant demographics

The mean age of the 9 patient participants was 66.6 years (SD = 6.3), with just over half being female (n = 5). All except one of the patient participants identified as white British, this patient participant identified as mixed race. All reported having at least a college-level education. Only two were currently employed, 1 full-time the other part-time. All others said that they had retired. Most of the sample (n = 6, 66.7%) reported their IBM severity as moderate, n = 1 (11.1%) reported it as mild, and n = 2 (22.2%) as severe.

The mean time since symptom onset was 10.9 (SD = 6.4) years with symptoms first occurring in the proximal lower limbs for most patient participants (n = 8, 88.9%). The mean time since IBM diagnosis was 4.6 (SD = 5.0) years. All patients were either diagnosed by clinico-pathologically defined IBM (n = 5, 55.6%), or clinically defined IBM (n = 4, 44.4%). Physician-reported patient severity was noted as: moderate (n = 4; 44.4%), mild (n = 2, 22.2%), severe (n = 2, 22.2%), and very severe (n = 1, 11.1%).

HCP demographics

Table 1 presents the demographic data of the physicians (Stage 1) and raters (Stages 3-5). At Stage 5, one of the raters only completed the video ratings at time point 1 therefore they were excluded from the Stage 5 analysis.

A total of 17 HCPs took part in this study, 14 were from the UK and 3 were from the US. The 3 US raters only took part in Stages 1 and 5 of the study. During Stage 1 all HCP participants were required to be physicians. For all other stages of the study, raters were a mix of physicians, nurses, and physiotherapists. The two most reported areas of specialization were neurology and physiotherapy. There was a range of qualifications reported including medical doctorates, PhDs, undergraduate/postgraduate qualifications, as well as those with nursing and physiotherapy qualifications. The HCPs across the stages had a wide range of years of experience treating IBM patients i.e., fewer than 5 to over 15 years. Most raters and physicians n = 16 (88.9%) selected the 1-250 option for a number of patients seen during their IBM career.

Qualitative Results

Stages 1 (physician) and 2 (patient participants) phone and face-to-face interviews, respectively, revealed that IBM has a substantial impact on patient participants' day-to-day lives, with a particular impact on physical function. The concepts identified can be broadly grouped as: physical upper limb, physical lower limb, oral, social, and emotional.

Physical function limitations related to either upper or lower limbs were some of the most frequently identified impacts of IBM. Upper limb difficulties focused on difficulties related to dressing, handwriting, fine motor skills, cutting food and using utensils, and hygiene (cleaning and washing themselves). Patient participants described how when dressing they had to select clothes that were easy to put on and did not involve zip or buttons, "*clothes that, that just go on very comfortably... and very easily, 'cause zips, buttons become a problem*." Patient handwriting had become more illegible as the condition worsened, "*Um, my handwriting is not as good as it used to be... but I can still write*." Whilst patients found it hard to complete fine motor skills such as opening jars and doors, "*it, has an effect on the fingers, so you know, grasping things. Turning, opening doors, jar tops*", or use utensils to cut food "*I drop things all the time. The knife drops, the fork drops-... the spoon drops, the milk gets spilled.*"

For lower limbs difficulties moving from sitting to standing, walking, climbing stairs, falling, and turning in bed were key areas identified. Walking was one of the most impacted areas with participants ranging from having minor difficulties to using a walking stick or frame, or being wheelchair bound "*Um, over time they become confined to a wheelchair with... You know - some of them can't ambulate on their own and others at all [which has a big impact]*." Similarly, as the condition worsened patients were unable to use stairs, "*haven't used the, uh, going upstairs since February (9 months)*" and would have difficulties moving from sitting to standing and for those who were most severe would need a hoist, *"[I need] a sling and a hoist*".

In reviewing the IBMFRS, all physicians said that the items were relevant and that the response options were suitable. Similarly, for patient participants, the majority felt the functional behaviors were relevant and important. However, 2 patient participants indicated that swallowing, handwriting, cutting food, and turning in bed were not relevant currently because their IBM was relatively mild.

The IBMFRS captured all core functional impacts identified during the qualitative interviews. Table 2 maps all the identified concepts to the items in the IBMFRS. The core functional concepts captured are swallowing, handwriting, cutting and handling utensils, fine motor skills, dressing, hygiene, turning in bed and adjusting covers, sit to stand, walking, and climbing stairs.

A few concepts were not captured in the IBMFRS: loss of hobbies and social life (n = 10), low mood and depression (n = 3), impact on driving (n = 3), impact on work (n = 3), and weight loss (n = 3); falls (n = 15).

As part of the discussion, some patient participants (n = 8) and physicians (n =7) also described what they would consider a meaningful change on the IBMFRS. Most (n=5) of the physicians indicated that a 1-category change on any item would represent a meaningful change in functional behaviors for the patient. This was also reflected by all 8 of the patient participants who indicated that a 1-category change on any item would be a meaningful change for them, *"Well, they all [change in response options] would be [meaningful], w- because they'd all have an impact on your life. I mean, for example, swallowing, the next one down, is occasional choking, which is not pleasant"*. Two of the 7 clinicians considered that a 2-category change on any item represented a meaningful change in functional behaviors.

Quantitative Results

The quantitative analysis demonstrated strong reliability within and between raters (Table 3). Inter-rater reliability from the face-to-face ratings and when against video was also very good, with ICCs of 0.847 and 0.971, respectively.[(Koo and Li 2016)] Intra-reliability using video ratings had an excellent ICC of 0.999. Equivalence between the different modes of administration, i.e., face-to-face versus phone, was very high, 0.941.

Overall, strata, separation, reliability of separation, and chi-square *p*-values were all in agreement with the expectations of the many-facets model. Specifically, the average correlation between a single rater and the rest of the raters R_c was .85, which in line with high ICCs reported for classical statistically inter-rater reliability, indicates high inter-rater agreement. For the Stage and Rater facets, the non-significant p-value for the chi-square test, low strata, separation, and reliability of separation, suggested that there are no observed differences between raters at each stage in terms of their severity. Conversely, for Patient facet, the significant p-value for chi-square test, high strata, separation, and reliability of separation, suggested that there are no observed differences between patients' severity levels.

DISCUSSION

Clinical outcome assessment measures are a valuable tool for assessing the symptom burden and impact of a disease on patients' health related quality of life.(Walton, Powers III et al. 2015) However, they require sufficient evidence supporting their reliability and validity prior to use in clinical trials.(Morrow, Ramdharry et al. 2013) Although there is some psychometric evidence of the IBMFRS(Jackson, Barohn et al. 2008, Morrow, Ramdharry et al. 2013, Sangha, Yao et al. 2021) there was a paucity of evidence demonstrating the instrument's content validity, inter-rater and intra-rater reliability and, what constitutes a meaningful change, all of which have been addressed in this current study.

The analysis of qualitative data from the patient and physician interviews revealed that IBM has a substantial impact on patient participants' lives, particularly with regards to functional impacts. The IBMFRS adequately captures all core functional impacts of IBM, which was corroborated by both patient participants and physicians when debriefing the measure. The concept of "falls" mentioned as important by physicians is not explicitly captured by an item in the IBMFRS; however, physicians considered that this would be included in rating the patient's walking ability. A few other concepts discussed by patient participants are not captured in the IBMFRS i.e., loss of hobbies and social life, low mood and depression, impact on driving, impact on work, and weight loss. However, these were typically infrequently discussed and were not functional impacts, and therefore would not be suitable for inclusion in a measure of function. They could be addressed using other scales designed to capture these and other related health domains if required.

When discussing meaningful change, most patient participants and physicians agreed that a 1-category change on any of the IBMFRS items, in any direction, would be meaningful. This highlights not only the debilitating nature of the disease, but the therapeutic potential of a treatment as even small changes would be considered meaningful to patients.

Furthermore, the inter- and intra-reliability of the IBMFRS was excellent, showing consistency between raters and stability over time, respectively. There was also excellent correlation between face-to-face and phone administration showing equivalence regarding the mode of administration. Additionally, the supportive Facet analysis further demonstrated agreement between all raters, as well as showing agreement between raters when compared to the gold standard rater.

Although originally derived from the ALS Functional Rating Scale, (Cedarbaum and Stambler 1997) this study demonstrates the relevance of the IBMFRS and its suitability for reliably assessing function in patients with IBM. However, given that the IBMFRS is a more holistic and global measure of physical function with a strong emphasis on upper and lower limb function, the IBMFRS might not be suitable for studies focusing on dysphagia as a key study endpoint (out of 10 items, only 1 item addresses dysphagia). Should researchers wish to investigate swallowing or respiratory function in greater detail, then other bulbar-specific measures such as the Eating Assessment Tool-10 (Belafsky, Mouadeb et al. 2008) or the modified oculobulbar facial respiratory score (Goyal, Araujo et al. 2017) should be considered.

Limitations

The study was undertaken using only UK patient participants; however cultural differences in the functional behaviors endorsed are not anticipated. Additionally, using the good practice guidance for the translation and cultural adaptation (Wild, Grove et al. 2005) of a measure would mitigate any issues when using the instrument cross-culturally. The study was also undertaken during the COVID-19 pandemic, therefore, only 9 patient participants could be recruited, and these patients took part in all

relevant stages of the study (Stage 2, 3, and 4). This was sufficient for the content assessment where 10 interviews were planned, although additional patient participants at different functional status would have been useful to determine reliability at the extremes of the disease. However, for Stages 3 and 4 (face-to-face and phone rating) 14 patient participants were originally planned. Despite this there was still a good range of severities and differences in abilities such that inter-rater reliability was arguably still robustly tested.

CONCLUSION

Overall, it is concluded that the IBMFRS is content valid in assessing the key functional impacts of IBM, is reliable both within and across raters, and there is equivalence of scores between face-to-face and phone modes of administration highlighting its feasibility for use within the context of clinical trials, with a 1-category change on any items being classed as meaningful by both patient and physicians.

References

Alexanderson, H. (2012). "Exercise in inflammatory myopathies, including inclusion body myositis." Current rheumatology reports **14**(3): 244-251.

Belafsky, P. C., et al. (2008). "Validity and reliability of the Eating Assessment Tool (EAT-10)." Annals of Otology, Rhinology & Laryngology **117**(12): 919-924.

Braun, V. and V. Clarke (2006). "Using thematic analysis in psychology." Qualitative research in psychology **3**(2): 77-101.

Callan, A., et al. (2017). "A systematic review and meta-analysis of prevalence studies of sporadic inclusion body myositis." Journal of neuromuscular diseases **4**(2): 127-137.

Cedarbaum, J. M. and N. Stambler (1997). "Performance of the amyotrophic lateral sclerosis functional rating scale (ALSFRS) in multicenter clinical trials." Journal of the neurological sciences **152**: s1-s9.

Dimachkie, M. M. and R. J. Barohn (2009). Idiopathic inflammatory myopathies. Immune-Mediated Neuromuscular Diseases, Karger Publishers. **26:** 126-146.

Dimachkie, M. M. and R. J. Barohn (2014). "Inclusion Body Myositis." Neurologic Clinics **32**(3): 629-646.

Goyal, N., et al. (2017). Feasibility and validation of modified oculobulbar facial respiratory score (mOBFRS) in amyotrophic lateral sclerosis (ALS) and sporadic inclusion body myositis (sIBM)(P1. 131), AAN Enterprises.

Jackson, C. E., et al. (2008). "Inclusion body myositis functional rating scale: a reliable and valid measure of disease severity." Muscle & nerve **37**(4): 473-476.

Koo, T. K. and M. Y. Li (2016). "A guideline of selecting and reporting intraclass correlation coefficients for reliability research." Journal of chiropractic medicine **15**(2): 155-163.

Landis, J. R. and G. G. Koch (1977). "The measurement of observer agreement for categorical data." biometrics: 159-174.

Lilleker, J., et al. (2017). "Cytosolic 5′-nucleotidase 1A autoantibody profile and clinical characteristics in inclusion body myositis." Annals of the rheumatic diseases **76**(5): 862-868.

Linacre, J. and B. Wright (1994). FACETS: Many-Facet Rasch Analysis, Chicago Il: MESA Press.

Machado, P., et al. (2013). "Update in inclusion body myositis." Current opinion in rheumatology **25**(6): 763.

McGraw, K. O. and S. P. Wong (1996). "Forming inferences about some intraclass correlation coefficients." Psychological methods **1**(1): 30.

Morrow, J., et al. (2013). "Rasch analysis of the IBMFRS." Muscle Nerve **48**: S2-S3.

Morrow, J. M., et al. (2013). "Rasch analysis of the IBMFRS." Muscle & nerve **48**: S2-S3.

Naddaf, E., et al. (2018). "Inclusion body myositis: update on pathogenesis and treatment." Neurotherapeutics **15**(4): 995-1005.

Naddaf, E., et al. (2022). "Survival and associated comorbidities in inclusion body myositis." Rheumatology **61**(5): 2016-2024. Erratum in: Rheumatology (Oxford). 2022 Apr 2009;: PMID: 34534271; PMCID: PMC39071572.

Qin, S., et al. (2019). "Assessing test–retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula." Quality of Life Research **28**(4): 1029-1033.

Revicki, D. A. (2018). Clinical Outcome Assessment Endpoints for Rare Diseases: Challenges and Methods for Clinical Trials, The International Society for CNS Clinical Trials and Methodology.

Rose, M. (2013). "188th ENMC international workshop: inclusion body myositis, 2–4 December 2011, Naarden, The Netherlands." Neuromuscular Disorders **23**(12): 1044-1055.

Sangha, G., et al. (2021). "Longitudinal observational study investigating outcome measures for clinical trials in inclusion body myositis." Journal of Neurology, Neurosurgery & Psychiatry **92**(8): 854-862.

SAS (2013). "Cary, NC: SAS Institute Inc."

Shelly, S., et al. (2021). "Epidemiology and natural history of inclusion body myositis: a 40-year population-based study." Neurology **96**(21): e2653-e2661.

Walton, M. K., et al. (2015). "Clinical outcome assessments: conceptual foundation—report of the ISPOR clinical outcomes assessment–emerging good practices for outcomes research task force." Value in Health **18**(6): 741-752.

Wild, D., et al. (2005). "Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation." Value in health **8**(2): 94-104.

Table 1. HCP Demographic Variables

Demographic Variables	Total N = 18[1]	Stage 1 N = 9[1]	Stage 3 and 4 N = 14	Stage 5 N = 13 [2]
Country				
UK	14	7	14	11
US	4	2	0	2
Qualifications, n (%)				
Undergraduate (BSc, BA, other)	8 (44.4)	4 (44.4)	6 (46.2)	7 (46.7)
Postgraduate (MSc, MA, other)	7 (38.9)	3 (33.3)	7 (53.8)	8 (53.3)
PhD	6 (33.3)	5 (55.6)	5 (38.5)	5 (33.3)
MD or equivalent	7 (38.9)	7 (77.8)	3 (23.1)	3 (20.0)
Other (total)	6 (33.3)	5 (55.6)	4 (30.8)	4 (26.7)
FRACP	2 (33.3)	2 (40.0)	0 (0.0)	0 (0.0)
MRCP UK, SCE Neurology, FRCP	1 (16.7)	1 (20.0)	1 (25.0)	1 (25.0)
MSc	1 (16.7)	1 (20.0)	1 (25.0)	1 (25.0)
RGN	1 (16.7)	0 (0.0)	1 (25.0)	1 (25.0)
Missing	1 (16.7)	1 (20.0)	1 (25.0)	1 (25.0)
Clinical specialty, n (%)[1]				
Neurologist	8 (44.4)	8 (88.9)	6 (42.9)	4 (30.8)
Rheumatologist	1 (5.6)	1 (11.1)	1 (7.1)	1 (7.7)
Physiotherapist	4 (22.2)	0 (0.0)	4 (28.6)	3 (23.1)
Other (total)	5 (27.8)	0 (0.0)	3 (21.4)	5 (38.5)
Clinical Nurse Specialist	2 (40.0)	0 (0.0)	2 (66.7)	2 (40.0)
Research Coordinator for Neurological Diseases	1 (20.0)	0 (0.0)	0 (0.0)	1 (20.0)
Research Nurse	1 (20.0)	0 (0.0)	1 (33.3)	1 (20.0)
Research Staff working with neurologists	1 (20.0)	0 (0.0)	0 (0.0)	1 (20.0)
Number of years treating IBM n (%)				
Fewer than 5 years	5 (27.8)	0 (0.0)	3 (21.4)	5 (38.5)
6-10 years	5 (27.8)	3 (33.3)	5 (35.7)	5 (38.5)
11-15 years	4 (22.2)	3 (33.3)	4 (28.6)	3 (23.1)
Over 15 years	4 (22.2)	3 (33.3)	2 (14.3)	0 (0.0)
Total IBM patients seen in career, n (%)				
1-250	16 (88.9)	7 (77.8)	13 (92.9)	12 (92.3)
251-500	1 (5.6)	1 (11.1)	1 (7.1)	1 (7.7)
501-750	1 (5.6)	1 (11.1)	0 (0.0)	0 (0.0)
751-1000	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Over 1001	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Previous experience of IBMFRS, n (%)				
Yes, in clinical practice	8 (44.4)	5 (55.6)	7 (50.0)	4 (30.8)
Yes, in research studies	8 (44.4)	6 (66.7)	5 (35.7)	5 (38.5)
No	9 (50.0)	4 (44.4)	8 (57.1)	8 (61.5)

Abbreviations: max = maximum; min = minimum; SD = standard deviation.

[1]Some physicians and raters selected more than one answer when asked about their highest qualification or area of clinical specialty.
At Stage 1, 1 US physician did not return their demographic form.
At Stage 5 rater demographics are based on those who completed time points 1 and 2.

Table 2. Mapping identified concepts to the IBMFRS

Qualitative Concept	IBMFRS	Note
Dressing	Item 5 - Dressing	Captured in IBMFRS
Fine motor skills	Item 4 - Fine motor tasks	Captured in IBMFRS
Handwriting	Item 2 - Handwriting	Captured in IBMFRS
Cutting food and handling utensils	Item 3 - Cutting and handling utensils	Captured in IBMFRS
Hygiene	Item 6 - Hygiene	Captured in IBMFRS
Walking	Item 9 - Walking	Captured in IBMFRS
Climbing stairs	Item 10 - Climbing stairs	Captured in IBMFRS
Sit to Stand	Item 8 - Sit to Stand	Captured in IBMFRS
Falls	X	This is not captured explicitly in the IBMFRS. Physicians discussed considering this when rating the patients' walking ability. No proposed changes to IBMFRS. If a count or greater details on falls is needed a separate measure should be used.
Turning in bed and adjusting covers	Item 7 - Turning in bed and adjusting covers	Captured in IBMFRS.
Loss of hobbies and social life	Not captured	Not captured. This would not be suitable for inclusion in the IBMFRS since it is not a functional behaviour.
Low mood and depression	Not captured	Not captured. This would not be suitable for inclusion in the IBMFRS since it is not a functional behaviour.
Swallowing	Item 1- Swallowing	Captured in IBMFRS.
Impact on Driving	Not captured	Not captured. The impact on driving was infrequently discussed in addition; this would not be suitable for inclusion in the IBMFRS as this is not a behaviour engaged in by all individuals. No proposed changes to IBMFRS.
Impact on work	Not captured	Not captured. The impact on work was infrequently discussed. This would also not be suitable for inclusion in the IBMFRS since this is not a behaviour engaged in by all individuals. No proposed change to IBMFRS.
Weight loss	Not captured	Not captured. This would not be suitable for inclusion in the IBMFRS since it is not a functional behaviour.

Table 3. Inter- and intra-rater reliability by Stage

Timepoint	$N_{Patients}$	N_{Raters}	$N_{ratings}$	Inter-rater Reliability	Intra-rater Reliability
Stage 3 (Face-to-Face)	9	14	49	0.847	---
Stage 3 (Face-to-Face) / Stage 4 (Telephone Rating)	9	14	48	---	0.941
Stage 5 (Time 1)	4	15	60	0.971	---
Stage 5 (Time 1) / Stage 5 (Time 2)	4	13	104	---	0.999