Published in final edited form as:

Qual Life Res. 2017 September ; 26(9): 2563–2572. doi:10.1007/s11136-017-1598-1.

Linking existing instruments to develop a continuum of care measure: accuracy comparison using function-related group classification

Chih-Ying Li¹, Sergio Romero^{2,3}, Kit N. Simpson^{4,5}, Heather S. Bonilha⁵, Annie N. Simpson^{4,5}, Ickpyo Hong¹, Craig A.VELOZO⁶

¹Division of Rehabilitation Sciences, School of Health Professions, University of Texas Medical Branch, Office 4.430, 301 University Boulevard, Galveston, TX 77555-1137, USA

²Department of Occupational Therapy, College of Public Health and Health Professions, University of Florida, Gainesville, FL 32610-0164, USA

³Department of Veterans Affairs, Center of Innovation on Disability and Rehabilitation Research (CINDRR), 151-B NF/SG VA Medical Center, 1601 SW Archer Rd., Gainesville, FL 32608, USA

⁴Department of Healthcare Leadership and Management, College of Health Professions, Medical University of South Carolina, Rm 410, 151-B Rutledge Ave, Charleston, SC 29425, USA

⁵Department of Health Sciences and Research, College of Health Professions, Medical University of South Carolina, 77 President St., MSC 700, Charleston, SC 29425, USA

⁶Division of Occupational Therapy, College of Health Professions, Medical University of South Carolina, Rm 330, 151-B Rutledge Avenue, MSC 962, Charleston, SC 29425-9620, USA

Abstract

Purposes—To compare measurement accuracy of test forms with varied number of items (13, 8, and 4 items) generated from the self-care physical function item bank composed of Functional Independence Measure (FIMTM) and the Minimum Data Set (MDS).

Methods—Retrospective data analysis of 2499 Veterans who completed both FIM and MDS within 6 days. We compared measurement accuracy between the converted FIM (FIMc) motor score generated from the MDS and the original FIM (FIMa) motor score (13 items) at: (a) individual-level using point differences, and (b) group-level using function-related group (FRG).

Results—The differences of mean FIMa and FIMc scores were between 0.05 and 1.07 points for all test forms. Over 81% of FIMc from MDS_13 were within 15 points of the FIMa. 81–90% of FRGs generated by the FIM short forms was identical to those generated by the FIMa for stroke, lower limb amputation, knee and hip replacement; and 59.9–90.5% by all MDS test forms. All MDS test forms had above 74% agreement with same or adjacent FMGs (ICC 0.65–0.91).

Chih-Ying Li, chili@utmb.edu.

Conflict of interest All authors declare that there is no conflict of interest with respect to publish this paper.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Electronic supplementary material The online version of this article (doi:10.1007/s11136-017-1598-1) contains supplementary material, which is available to authorized users.

Conclusions—The accuracy is dependent on the comparison level (i.e., individual or group), length of the test and which FRG is used. Our results partially support using existing instruments—without decreasing the number of the items—to generate a continuum of care measurement.

Keywords

Continuity of patient care; Outcome assessment (health care); Activities of daily living; Veterans; Classification

Introduction

The need for a functional measure across settings has resulted in efforts to develop a new, uniform, and standardized instrument. The Functional Assessment Standardized Items (FASI), developed by the Centers for Medicare & Medicaid Services (CMS), is the latest example of an instrument of this kind. CMS expects FASI to provide comparable documentation across the continuum of care, from acute to post-acute rehabilitation settings [1, 2]. However, adopting a new instrument could incur undesirable consequences such as considerable amounts of resources and cost prior to its development, and following reliability/validity testing, new tool's administration and training, or even further mandated reimbursement system change. There is a further major disadvantage of adopting a new instrument in the Veterans Administration (VA) healthcare system. The VA is well known and revered for its longitudinal database of medical records, including functional assessments. To adopt a new tool would essentially require abandoning the extensive retrospective functional outcome data and be unable to compare back with the existing outcome data unless certain linking method is used. This is not a trivial disadvantage as longitudinal medical, and patient care database is rare and enables unique research opportunities.

To address the issue that different instruments are currently used across post-acute settings, linking existing instruments into an item bank is an alternative approach. Particularly, linking can be a useful strategy to develop a continuum of care measurement in the VA settings. There is fairly strong evidence in the healthcare literature supporting the method of linking [3-6]. The item response theory (IRT)-based linking method places test items from different assessments that measure the same underlying construct on a common item bank. Item bank enables scores of different assessments to be converted and compared. Linking existing instruments to develop an item bank allows practitioners to continue using the validated instruments that are in use in their practice.

To demonstrate the above proposed idea, we used IRT-based method to create an item bank composed of the Functional Independence Measure (FIMTM), used in the inpatient rehabilitation facility (IRF), and the Minimum Data Set (MDS), used in the Community Living Center (CLC) in the VA system [7]. This self-care physical function item bank had a total of 26 items composed of 13 FIM and 13 MDS motor items, which have been examined for its item-level psychometric properties [7, 8]. In addition to the item bank, the IRT method could also allow short forms (SFs) development with the advantages of improving assessment efficiency and reducing assessment burden for both practitioners and patients.

We developed several SFs from this FIM-MDS item bank, including FIM_4- and 8-item, and MDS_4- and 8-item. All SFs were developed based on del Toro et al.'s [9] procedures. First, we excluded items with residual correlations higher than or equal to ± 0.2 to minimize item redundancy. Second, we selected items at least 2 standard errors apart from full range of estimates. Lastly, we chose items with item discrimination closest to 1 (i.e., best fit the Rasch model) [9]. Our research team has previously evaluated the measurement precision of all developed SFs [8]. The current study aims to evaluate measurement accuracy of the developed instruments. Accuracy was evaluated based on whether the converted scores from different instruments could classify patients into the same disability level as the original scores from the original instrument. If converted scores generated from the existing linked instrument could provide similar measurement accuracy as the original assessment, then the hypothesis of developing a continuum of measurement using existing instruments could be supported.

We used the function-related groups (FRGs), a form of case mix groups (CMGs), as a way to compare measurement accuracy. The FRGs are the basis for the IRF prospective payment system (PPS) [9]. Stineman and co-authors developed and validate the FRG algorithms to predict the costs of treating Medicare patients [10-12]. The FRG algorithms used the FIM physical functioning (13 items) and cognitive (5 items) scores, along with patients' age at admission to the IRFs. Based on the rehabilitation impairment classification, patients are classified into one of 20 diverse diagnoses [12]. Each diagnosis has a specific FRG algorithm resulting in a different number of FRG categories. Patients assigned to different FRG categories are expected to have different rehabilitation outcomes and total healthcare costs [12].

The FRGs classification system provides a pragmatic method to examine measurement accuracy at group level for the "converted" FIM score (i.e., FIM scores generated by different sets of MDS items from the item bank). We compare the scores derived from the original FIM and the different test forms to investigate whether the converted FIM scores could classify the same patient into the same classification level. We use the 4- and 8-item SFs generated from the item bank to produce FIM converted scores, and used the converted scores to assign FRGs. We hypothesized that due to similar inherent amount of errors of the performance-based instruments, the FRG generated from different sets of items from the item bank will be comparable. In other words, the SFs generated from either FIM items or the MDS items are hypothesized to generate similar FRGs categories for Veterans as those generated from the original FIM. At the group level, we hypothesized the similarity of accuracy to be at least 80% of exact agreement and one category apart of the FRGs generated by the actual and the converted score. At the individual level, the similarity of accuracy was hypothesized to be at least 80% of the converted scores fell within 15-point difference to the actual score. This study is the first step needed to provide evidence to support the feasibility/accuracy of using converted scores as a continuum of care measurement in rehabilitation research and clinical practice.

Methods

Ethical disclosure

This study was deemed Not Human Subjects Research by the Office of Research Integrity at the Medical University of South Carolina and approved by Health Center Institutional Review Board at the University of Florida (IRB Project # 440-2011) and the Veterans Health Service Human Research Protection Program (HRPP) at the North Florida/South Georgia Veterans Health System.

Participants

We used a retrospective dataset from both the VA Physical Medicine and Rehabilitation Service (10P4) and VA Geriatric and Extended Care Operations (10NC4) during October 2008 to September 2010. Stratified sampling was used to extract study sample based on four diagnoses: stroke, lower limb amputation, hip replacement, and knee replacement. These four diagnoses were chosen based on distinguishable diagnostic features used in FRG classifications (i.e., it is rare for the same patient to have more than one out of these four diagnoses at the same time). The only inclusion criterion was that participant completed both FIM and MDS without missing any items within 6 days. FIM motor score is the required component for all used FRG algorithms in this study. FIM motor scores, FIM cognitive scores and age of each Veteran were used to generate FRG, except for the diagnosis of lower limb amputation (Supplementary Figs. 1-4). The ability estimate based on the original FIM_13 was considered the “gold standard,” which was referred to as the FIM actual score (FIMa).

Measurement accuracy

MDS_13-item, MDS_8-item, and MDS_4-item scores were converted to the FIM_13 motor scores (FIM converted, FIMc). This conversion is based on the common metric feature of the IRT-based item bank. We examined measurement accuracy of all test forms at two levels: the individual and the group levels.

Measurement accuracy at individual-level—At the individual level, we used Wilcoxon signed rank-sum test to compare medians of two distributions between FIMa and FIMc since Kolmogorov–Smirnov testing showed that the scores were not normally distributed. We further calculated the percentage of converted scores that were within 5-, 10-, and 15-point difference of the actual scores (range of FIM motor score is 13–91 points). Point difference was defined as the absolute value calculated between the actual FIM (FIMa) and the converted FIM (FIMc). We also calculated percentage of converted scores within 5- and 10-point differences between FIMa and FIMc. Spearman’s rank correlation coefficient was calculated between the FIMa and FIMc for all test forms. The absolute value of correlation coefficient within a range of 0.90–1.00 was considered as very high, 0.70–0.90 was high, 0.50–0.70 was moderate, 0.30–0.50 was low, and 0.00–0.30 was negligible [13]. Intraclass correlations coefficients (ICC) were calculated between original FIM_13 (FIMa) and FIMc generated from all other test forms. We used a two-way mixed method to calculate absolute agreement (ICC(3,*k*)) with an assumption that reliability was calculated by an average of the *k* raters’ measurements. ICC values less than .40 were classified as poor,

between .40 and .59 as fair, between .60 and .74 as good, and between .75 and 1.0 as excellent [14]. A value of $\alpha < 0.05$ was considered statistically significant.

Measurement accuracy at group-level—At group level, we compared FRG classifications generated from each test form (FRGc) to the “actual” FRG by the FIMa (FRGa). We used FIMa and FIMc to classify each individual into one of the FRGs for each diagnosis and then calculated the percentage of agreement between FRGa and FRGc. Three FRG algorithms in total were used (knee replacement and hip replacement had the same algorithm). We only replaced FIM-motor score in all comparisons. We calculated the percentage of FRGa/FRGc falling into the same FRG category (exact agreement), one category apart (± 1 level), two categories apart (± 2 levels), and greater than two categories apart (± 3 to ± 7 levels).

We used weighted kappa to examine agreement strength of FRGa and FRGc, with kappa and McNemar’s test to provide a 2×2 table for the lower limb amputation FRG calculation due to its dichotomous FRG feature. A weighted kappa statistic for categorical data ranging from 0.21 to 0.40 demonstrates a fair strength of agreement, from 0.41 to 0.60 represents a moderate agreement, and from 0.61 to 0.80 indicates a substantial agreement [15]. We also calculated a two-way mixed ICC between the FRGa and FRGc for all test forms across the four diagnoses. In addition, the coefficient of variation of method error (CV_{ME}) was calculated to cross-validate kappa and ICC results. CV_{ME} is an indication of the amount of variation in the difference scores between two sets of repeated scores with the advantage for the data lack of variation and is independent of ICC [15]. The criteria of $CV_{ME} < 10$ represents reliable repeated scores of the test [16].

Results

Demographics

A total 2499 Veterans were included in this study after removing one person with missing age. To comply with the HIPAA requirement, 63 (2.5%) patients older than 89 years old (y/o) formed one group labeled as 90 y/o (Table 1). Mean age in the sample was 67.1 y/o ($SD = 11.2$) with a range from 19 to 90 y/o. The majority of the sample was male (96.2%), white (65.5%), married (42.5%), and resided at the acute unit (44.5%) or at home (20.1%) prior to their admission to the hospital. The individuals were diagnosed of stroke ($n = 1065$, 42.6%), lower limb amputation ($n = 472$, 18.9%), knee replacement ($n = 568$, 22.7%), and hip replacement ($n = 394$, 15.8%) (Table 1). The average length of days between the administrations of the FIM and the MDS was 3.2 days ($SD = 2.1$) (Table 1).

Correlations between FIMa and FIMc from varied test forms

The Spearman’s correlations with FIM_13, for MDS_13-, 8-, and 4-item, were 0.81, 0.78, and 0.75, respectively (Table 2). The correlation strengths between FIMa and FIMc generated from all MDS test forms were all significant and very high (p values all < 0.05). The converted scores generated from all MDS test forms had excellent ICCs with the FIM_13 scores (range 0.86–0.89) (Table 2).

Accuracy comparisons at individual level: point difference

The FIMa and FIMc generated from each test form all had negatively skewed distributions, indicating that individuals tended to have high scores (better self-care physical function). Wilcoxon signed rank-sum tests showed a significant difference in median score distribution between the FIMa and FIMc, regardless of which test form was compared (all p values <0.0001) (Table 2). FIM_8 and FIM_4 SFs had 100/98% of FIMc within 15 points of the FIMa, 95–99% of the scores within 10 points, and 78–92% of the scores within 5 points (Table 2). All MDS test forms (13, 8, and 4 items) had 72–81% of FIMc within 15 points of the FIMa, 56–65% of the scores within 10 points, and 31–39% of the scores within 5 points (Table 2). The distributions of absolute point differences of all test forms were positively skewed, indicating low point differences across all test forms (Supplementary Fig. 5a-e).

Accuracy comparisons at group level: FRG

Table 3 presents the percentage of FRGc that were exact, within 1 or more deviated classifications of the FRGa. Overall, both FIM SFs had higher FRGc agreement than all MDS test forms. For SFs, the FIM_8-item SFs had the highest exact agreement (80.66–97.97%) and MDS_4-item SFs had the lowest (59.91–80.93%) across all four diagnoses (Table 3).

For the stroke FRG, the range of exact agreement of FRGc for all test forms was between 59.91–85.16%, agreement ± 1 category was 74.46–95.67%, and agreement ± 2 categories was 80.75–97.74% (Table 3). For lower limb amputation, since this diagnosis only had two FRG categories, thus, the range of exact agreement of amputation FRGc for all test forms was between 80.93–95.34% for all test forms. Both 4- and 8-item FIM SFs had above 90% exact agreement. MDS_13, MDS_4 SF, and MDS_8 SF had above 80.93% exact agreement (Table 3). For knee replacement, the range of exact agreement of FRGc for all test forms was between 78.35–97.71%, agreement ± 1 category was 92.26–98.60%, and agreement ± 2 categories was 94.9–99.83%. FIM_8, FIM_4, and MDS_13 all had above 90% exact agreement while MDS_8 and MDS_4 both had 78.35% exact agreement (Table 3). For hip replacement, the range of exact agreement of FRGc for all test forms was between 69.80–97.97%, agreement ± 1 category was 84.52–98.98%, and agreement ± 2 categories was 92.89–100%. Overall, MDS_13-item had the highest exact agreement for knee replacement FRG and lowest exact agreement for stroke FRG. For all test forms, the knee replacement and hip were the two diagnoses that had the highest exact FRG agreement, while stroke had the lowest (Table 3).

Agreement strength of FIMc is presented in Table 4. Overall, within each test form, strength of agreement decreased when the number of items decreased. For stroke, agreement strength ranged from 0.69 to 0.73 and excellent ICCs (0.88–0.91), indicating strong agreement (Table 4). For lower limb amputation and hip replacement, all weighted kappa and kappa results were significant and the MDS SFs showed fair to good agreement (0.34–0.55) and good to excellent ICCs (0.65–0.80) (Table 4). However, for knee replacement, all weighted kappa statistics indicated significant but very weak agreement (0.09–0.17) and poor to fair ICCs (0.22–0.40) (Table 4). We recognized that kappa and ICC statistics could only provide accurate test results for the diagnoses with adequate variability. Thus, as a validation

concern, CV_{ME} was used to examine agreement strength for the knee replacement FRGs due to its insufficient variability of the responses. The results of CV_{ME} confirmed that only the knee replacement FRG had $CV_{ME} < 10$ (2.47, 2.69 and 4.13 for MDS_13, MDS_8, and MDS_4, respectively), indicating high reliability between the FRGa and FRGc for this diagnosis. In contrast, the stroke FRG had CV_{ME} between 20.92 and 22.63 (>10), indicating less reliability (i.e., more variations) between the FRGa and FRGc for stroke. For hip replacement, CV_{ME} were between 11.37 and 12.90, marginally above the criteria of 10, indicating good to moderate reliability (Table 5). Please note that the CV_{ME} for the amputation FRG were also below 10 (0.48–0.54), but this was due to its dichotomous FRG nature.

Discussion

The overarching goal of this study was to test the hypothesis of using existing instruments to develop an accurate continuum of care measurement. Comparing measurement accuracy of the converted score from the varied test forms is the first step to test this hypothesis. The study showed mixed results and partially supported our hypothesis regarding measurement accuracy when comparing the converted FIM scores (FIMc) with the actual FIM scores (FIMa) from varied test forms. Overall, the MDS_13 converted score had better measurement accuracy than converted scores generated from the 4-item and 8-item MDS short forms. Less number of items resulted in less accuracy since more random errors may occur [17]. The accuracy of SFs could be improved if errors could be minimized during instrumental assessment, SF development and score-translational processes.

Our study suggested converted scores should be used based on specific clinical purposes. For instance, the converted FIM score generated from the MDS could be useful for the clinicians if the purpose is to obtain a general sense of patients' function at IRF admission (when the patient is transferred from CLC). However, when using the individual converted scores to follow-up patients' progression over time, adjusting potential errors seems to be necessary. One solution is to use a plausible range of converted FIM values such as 95% confidence interval to minimize the error and also reduce the need to introduce arbitrary cutoff points. However, we did not intend to use range score of converted FIM values in this study.

The measurement accuracy findings at the individual level showed a significant difference of distributions between FIMa and FIMc for all test forms, even though we assumed the Wilcoxon signed rank-sum test may provide overly sensitive indications of significance due to large sample size, and therefore the skewness of the distributions could not be detected. Individual-level comparisons showed averaged around 10-point difference between the original FIM and the converted scores from all the MDS test forms. Above 81% of the converted scores from MDS_13 was within 15 points, indicating a high proportion (more than 2/3) of the converted MDS_13 scores was within the FIM motor actual score, which was above hypothesized accuracy criteria (80%). Compared to Wang et al.'s results [6], our study showed similar but slightly better results with 38.7% (vs 33.7%) within 5 points and 64.7% (vs 56.9%) within 10 points between the FIMa and the FIMc generated from MDS_13-item. At the individual level by using point difference, the FIMc generated from

all MDS test forms (MDS_13, MDS_8 and MDS_4) had excellent ICC and high correlations with the FIMa (FIM_13). The MDS_13-item has better measurement accuracy than the two MDS SFs at the individual level. While the 8-item and 4-item MDS SFs had 6–9% less of point difference within 15 points compared to the MDS_13-item.

The results showed all MDS test forms had 60–91% exact agreement and 79–94% within ± 1 FRG difference across four diagnoses. Stroke had the worst, while knee replacement had the highest FRG exact agreement. At the group level, three diagnoses (except stroke) had above hypothesized 80% agreement within one FRG apart. This is likely due to the greater varied functional levels in patients with stroke compared to those with knee replacement. In addition, the knee replacement FRG algorithm had fewer FRG categories compared to stroke (7 vs 9). However, it is important to note that traditional agreement testing method of using kappa (or weighted kappa) and ICC may mislead the results in a particular scenario when minimal variability was shown in the responses since higher agreement contradictorily resulted in less data variability. This happened for the knee replacement FRG in this study. Thus, it is important to check for adequate variability in the data prior to using kappa statistics to avoid misinterpretation. We suggested using CV_{ME} to cross-validate the weighted kappa results for the future study of this kind.

Previous crosswalk-validation studies used intraclass correlation coefficients (ICC) or Cohen's effect size at group level to compare score translatability between instruments [6, 18–23]. Our study showed a similar but slightly wider range of ICC as Klooster et al.'s study [23] between observed and translated scores (0.67–0.91 vs 0.72–0.82, excluding knee replacement). While most studies applied linking methods at the group-level, it is notable that the individual converted score may not be reliable as expected [6, 18, 20, 23, 24]. For instance, Holzner and colleagues [20] found the confidence intervals of translated scores for individual subjects were very large, limiting precision of individual scores after score conversion. Fischer et al. [24] also found similar results. Askew et al. [18] recommended that individual scores derived from crosswalks should be used for group-level analysis when considering inherent conversion errors. Our results were consistent with those findings. We suggested that score conversion had better measurement accuracy at the group level compared to the individual level. Along with our previous measurement precision study [8], we found MDS_13 converted scores had similar precision and accuracy compared to using the original FIM_13 at the group level. However, the MDS short forms only had comparable measurement precision but not accuracy as original FIM_13.

Combining existing instruments rather than generating new items to construct a new, single and standardized continuum of care measurement has several advantages for healthcare policy makers, researchers, clinicians and patients. Our study addressed the biggest concern of the linked existing instrument by comparing measurement accuracy of converted scores at the individual and group levels. With an expectation that inherent measurement errors of the FIM and the conversion errors during the conversion process both, considering one FRG category apart when using converted score at the group level is reasonable. We suggest using the conversion MDS_13 score (at the group level) for knee replacement, hip replacement, and lower limb amputation only (since these three diagnoses had above 80% agreements). The conversion system may not be appropriate for the stroke population since it did not

achieve 80% agreement. Future study will be needed to address this concern. Currently, no study investigates the impact of error level between using a single instrument and using linked instruments to convert the scores across the continuum of care. If the error levels are similar between these two approaches, then it is worth re-evaluating the need to develop a single new instrument.

Study limitations

Since stability of a patient's response is crucial to obtaining reliable measurement accuracy, the major limitation in this study was that we assumed patients' ability did not change within 6 days, which might not be true. Secondly, this study used retrospective data; thus, the data were collected without an intent to answer the proposed research hypotheses. Also, the data used in this study may be subject to inherent errors from all possible uncontrollable sources during the data collection process and also due to the fact that we used a random sample from a population. In addition, we did not compare similarities or inconsistencies of short-form development methods based on different IRT models; however, we assumed the variations of measurement accuracy of varied short-form methods may be minimum. Also, this study may not be generalizable to the other populations beyond US Veterans since post-acute mechanism and practices may differ. In addition, this study did not use an estimated range of the converted FIM values to reduce potential conversion errors. Lastly, while we used an 80% exact matching success rate to consistently compare FRG agreement across diagnoses, we recognize this cutoff is arbitrary. Hence, we suggest the exact cutoff criteria for similar future studies should differ based on the intended research question and the number of FRG categories.

Conclusion

This study partially supports the notion of using existing instruments to develop an accurate continuum of care measurement by showing MDS_13 could generate comparable scores (as FIM test forms) with equivalent measurement accuracy at individual and group levels. However, the magnitude of measurement errors each factor contributes to based on different measurement method (e.g., new single instrument vs linking instruments) remains unknown. This study serves as the first step to explore the feasibility of using converted scores as a continuum of care measurement in healthcare professions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This study was funded by the Department of Veterans Affairs, Health Services Research and Development, Grant IR 11-223-1, "Item Banking across the Continuum of Care," from North Florida/South Georgia Veterans Health System, Center of Innovation on Disability and Rehabilitation Research (CINDRR).

References

1. Centers for Medicare & Medicaid Services. (2012). U.S. Department of Health and Human Services. Report to congress: post acute care payment reform demonstration (PAC-PRD), January, 2012. Retrieved May 22, 2013, from http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Reports/downloads/Flood_PACPRD_RTC_CMS_Report_Jan_2012.pdf.
2. Centers for Medicare and Medicaid Services. (2015). IMPACT Act of 2014 & Cross Setting Measures. Retrieved June, 17, 2015, from <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/IMPACT-Act-of-2014-and-Cross-Setting-Measures.html>.
3. Fisher WP Jr., Harvey RF, Taylor P, Kilgore KM, & Kelly CK (1995). Rehabits: A common language of functional assessment. *Archives of Physical Medicine and Rehabilitation*, 76(2), 113–122. [PubMed: 7848069]
4. McHorney CA (2002). Use of item response theory to link three modules of functional status items from the Asset and Health Dynamics among the Oldest Old Study. *Archives of Physical Medicine and Rehabilitation*, 83, 383–394. [PubMed: 11887121]
5. Velozo CA, Byers KL, Wang YC, & Joseph BR (2007). Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set. *Journal of Rehabilitation Research and Development*, 44(3), 467–478. [PubMed: 18247243]
6. Wang YC, Byers KL, & Velozo CA (2008). Validation of FIM™-MDS crosswalk conversion algorithm. *Journal of Rehabilitation Research and Development*, 45(7), 1065–1076. [PubMed: 19165695]
7. Li CY, Romero R, Bonilha H, Simpson K, Simpson A, Hong I, et al. (2016a). Linking existing instruments to develop an activity of daily living item bank. *Evaluation and the Health Professions*. doi:10.1177/0163278716676873.
8. Li CY, Romero S, Simpson A, Simpson K, Bonilha H, Hong I, & Velozo C (2016b). Comparisons of functional independence measure-minimum data set short forms (manuscript submitted for publication).
9. del Toro CM, Bislick LP, Comer M, Velozo C, Romero S, Gonzalez Rothi LJ, et al. (2011). Development of a short form of the Boston naming test for individuals with aphasia. *J Speech Lang Hear Res*, 54(4), 1089–1100. doi:10.1044/1092-4388. [PubMed: 21173387]
10. Stineman MG (1995). Case-mix measurement in medical rehabilitation. *Archives of Physical Medicine and Rehabilitation*, 76(12), 1163–1170. [PubMed: 8540795]
11. Stineman MG, Escarce JJ, Goin JE, Hamilton BB, Granger CV, & Williams SV (1994). A case mix classification system for medical rehabilitation. *Medical Care*, 32, 366–379. [PubMed: 8139301]
12. Stineman MG, Tassoni CJ, Escarce JJ, Goin JE, Granger CV, Fiedler RC, et al. (1997). Development of function related groups version 2.0: a classification system for medical rehabilitation. *Health Services Research*, 32, 529–548. [PubMed: 9327817]
13. Mukaka MM (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3), 69–71. [PubMed: 23638278]
14. Hallgren KA (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials Quantitative Methods Psychology*, 8(1), 23–34.
15. Landis JR, & Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. [PubMed: 843571]
16. Portney LG, & Watkins MP (2009). Chapter 26. Statistical measures of reliability. In Leslie G 3rd (Ed.), *Foundations of clinical research: Applications to practice* (pp. 609–612). Upper Saddle River: Prentice Hall. ISBN 0-13-171640-9.
17. Boundless. (2016). Accuracy, precision, and error. Boundless Chemistry Boundless, 12 Aug 2016. Retrieved Mar 06, 2017, from <https://www.boundless.com/chemistry/textbooks/boundless-chemistry-textbook/introduction-to-chemistry-1/measurement-uncertainty-30/accuracy-precision-and-error-190-3706/>.

18. Askew RL, Kim J, Chung H, Cook KF, Johnson KL, & Amtmann D (2013). Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. *Quality of Life Research*, 22, 2769–2776. [PubMed: 23539469]
19. Bjorner JB, Kosinski M, & Ware JE Jr. (2003). Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Quality of Life Research*, 12, 981–1002. [PubMed: 14651417]
20. Holzner B, Bode RK, Hahn EA, Cella D, Kopp M, Sperner-Unterweger B, et al. (2006). Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research. *European Journal of Cancer*, 42(18), 3169–3177. [PubMed: 17045472]
21. Orlando M, Sherbourne CD, & Thissen D (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, 12, 354–359. [PubMed: 11021160]
22. Oude Voshaar MA, Ten Klooster PM, Taal E, Wolfe F, Vonkeman H, Glas CA, et al. (2014). Linking physical function outcomes in rheumatology: performance of a crosswalk for converting Health Assessment Questionnaire scores to short form 36 physical functioning scale scores. *Arthritis Care & Research*, 66(11), 1754–1758. [PubMed: 24757106]
23. Ten Klooster PM, Oude Voshaar MA, Gandek B, Rose M, Bjorner JB, Taal E, et al. (2013). Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment Questionnaire disability index in rheumatoid arthritis. *Health Quality of Life Outcomes*, 11, 199. doi:10.1186/1477-7525-11-199. [PubMed: 24229416]
24. Fischer HF, Tritt K, Klapp BF, & Fliege H (2011). How to compare scores from different depression scales: Equating the patient health questionnaire (PHQ) and the ICD-10-symptom rating (ISR) using item response theory. *International Journal of Methods in Psychiatric Research*, 20, 203–214. [PubMed: 22021205]

Table 1

Demographic characteristics of participants in this study (*n* = 2499)

Variables	Community-dwelling veterans (<i>n</i> = 2499)	
	Mean	SD
Age (range 19–90 y/o)	67.1	11.3
Averaged number of days since onset	155.1	1083.9
Variables	Community-dwelling veterans (<i>n</i> = 2499)	
	Number	%
Gender		
Male	2376	95.1
Female	93	3.7
Missing	30	1.2
Ethnicity		
White	1575	63.0
Black	582	23.3
Asian	8	0.3
Native American	11	0.4
Hispanic	129	5.2
Other	98	3.9
Missing	96	3.8
Diagnoses		
Stroke	1065	42.6
Lower limb amputee	472	18.9
Knee replacement	568	22.7
Hip replacement	394	15.8
Marital status		
Single	306	12.2
Married	1063	42.5
Widowed	160	6.4
Separated	89	3.6
Divorced	779	31.2

Variables	Community-dwelling veterans (<i>n</i> = 2499)	
	Mean	SD
Missing	102	4.1
Admission condition		
Initial rehabilitation	2362	94.5
Short stay evaluation	60	2.4
Readmission	15	0.6
Unplanned discharge without assessment	2	0.08
Continuing rehabilitation	56	2.2
Missing	4	0.2
Pre-living setting		
Home	502	20.1
Board and care	9	0.4
Transitional living	8	0.3
Intermediate care	12	0.5
Skilled nursing facility	143	5.7
Acute unit of own facility	1113	44.5
Acute unit of another facility	313	12.5
Chronic hospital	1	0.04
Rehabilitation facility	41	1.6
Other	11	0.4
Alternate level of care unit	1	0.04
Subacute unit	3	0.1
Assisted living residence	5	0.2
Missing	337	13.5
Days between administrations of FIM and MDS (range = 0–6)	Mean = 3.2	(<i>SD</i> = 2.1)

Table 2

Summary of FIM_13 raw scores and converted FIM raw score generated from varied test forms (FIM_8, FIM_4, MDS_13, MDS_8, MDS_4) ($n = 2500$)

	Median	Variance	Wilcoxon signed rank (compared with FIM_13)	Correlation (compared with FIM_13)	Intraclass correlation coefficients (ICC)	Point difference (FIM_13- converted FIM)		
						5 points (%)	10 points (%)	15 points (%)
FIM_13	77.00	523.71						
FIM_8	76.00	522.07	$p < 0.0001^*$	0.99 [*]	0.99 (excellent)	92.20	99.04	99.92
FIM_4	75.00	563.97	$p < 0.0001^*$	0.97 [*]	0.98 (excellent)	78.27	94.56	97.92
MDS_13	73.00	462.42	$p < 0.0001^*$	0.81 [*]	0.89 (excellent)	38.70	64.71	81.11
MDS_8	71.00	501.29	$p < 0.0001^*$	0.78 [*]	0.88 (excellent)	31.05	56.86	75.19
MDS_4	71.00	482.16	$p < 0.0001^*$	0.75 [*]	0.86 (excellent)	31.21	56.14	72.27

* Significant difference <0.05

Table 3

FRG classification difference between FIM_13 and other test forms across four diagnoses (stroke, lower limb amputation, knee replacement, and hip replacement)

FRG Difference (FIM_13 -)	FIM_8SF		FIM_4SF		MDS_13		MDS_8SF		MDS_4SF	
	Percent, N (%)	Cumulative, N (%)	Percent, N (%)	Cumulative, N (%)	Percent, N (%)	Cumulative, N (%)	Percent, N (%)	Cumulative, N (%)	Percent, N (%)	Cumulative, N (%)
Stroke (N = 1065)										
0	85.16 (907)	85.16 (907)	80.66 (859)	80.66 (859)	64.13 (683)	64.13 (683)	62.82 (669)	62.82 (669)	59.91 (638)	59.91 (638)
±1	10.51 (112)	95.67 (1019)	11.46 (122)	92.12 (981)	14.74 (157)	78.87 (840)	15.03 (160)	77.85 (829)	14.55 (155)	74.46 (793)
±2	2.07 (22)	97.74 (1041)	3.85 (41)	95.97 (1022)	6.38 (68)	85.25 (908)	7.14 (76)	84.99 (905)	6.29 (67)	80.75 (860)
±3	2.25 (24)	100 (1065)	3.66 (39)	99.63 (1061)	7.79 (83)	93.04 (991)	8.17 (87)	93.16 (992)	11.08 (118)	91.83 (978)
±4			0.28 (3)	99.91 (1064)	3.01 (32)	96.05 (1023)	3.85 (41)	97.01 (1033)	5.17 (55)	97 (1033)
±5					2.63 (28)	98.68 (1051)	1.88 (20)	98.89 (1053)	1.79 (19)	98.79 (1052)
±6					0.84 (9)	99.52 (1060)	0.75 (8)	99.64 (1061)	0.85 (9)	99.64 (1061)
±7			0.09 (1)	100 (1065)	0.47 (5)	100 (1065)	0.38 (4)	100 (1065)	0.38 (4)	100 (1065)
Lower limb amputation (n = 472)										
0	95.34 (450)	95.34 (450)	92.37 (436)	92.37 (436)	82.42 (389)	82.42 (389)	82.84 (391)	82.84 (391)	80.93 (382)	80.93 (382)
±1	4.66 (22)	100 (472)	7.63 (36)	100 (472)	17.59 (83)	100 (472)	17.16 (81)	100 (472)	19.07 (90)	100 (472)
Knee replacement (n = 568)										
0	97.54 (554)	97.54 (554)	97.71 (555)	97.71 (555)	90.49 (514)	90.49 (514)	78.35 (445)	78.35 (445)	78.35 (445)	78.35 (445)
±1	1.06 (6)	98.60 (560)	0.53 (3)	98.24 (558)	3.88 (22)	94.37 (536)	15.14 (86)	93.49 (531)	13.91 (79)	92.26 (524)
±2	1.23 (7)	99.83 (567)	1.05 (6)	99.29 (564)	2.46 (14)	96.83 (550)	2.99 (17)	96.48 (548)	2.64 (15)	94.9 (539)
±3	0.18 (1)	100 (568)	0.18 (1)	99.47 (565)	0.53 (3)	97.36 (553)	0.53 (3)	97.01 (551)	0.53 (3)	95.43 (542)
±4			0.36 (1)	99.83 (567)	2.29 (13)	99.65 (566)	2.64 (15)	99.65 (566)	3.7 (21)	99.13 (563)
±5			0.18 (1)	100 (568)	0.35 (2)	100 (568)	0.35 (2)	100 (568)	0.88 (5)	100 (568)
Hip replacement (n = 394)										
0	97.97 (386)	97.97 (386)	94.67 (373)	94.67 (373)	85.28 (336)	85.28 (336)	71.57 (282)	71.57 (282)	69.80 (275)	69.80 (275)
±1	1.01 (4)	98.98 (390)	1.78 (7)	96.45 (380)	6.09 (24)	91.37 (360)	15.73 (62)	87.3 (344)	14.72 (58)	84.52 (333)
±2	1.01 (4)	100 (394)	3.04 (12)	99.49 (392)	4.06 (16)	95.43 (376)	7.1 (28)	94.4 (372)	8.37 (33)	92.89 (366)
±3			0.5 (2)	100 (394)	2.03 (8)	97.46 (384)	3.04 (12)	97.44 (384)	3.3 (13)	96.19 (379)
±4					1.78 (7)	99.24 (391)	1.77 (7)	99.21 (391)	2.54 (10)	98.73 (389)

FRG Difference (FIM_13 -)	FIM_8SF			FIM_4SF			MDS_13			MDS_8SF			MDS_4SF		
	Percent, N (%)	Cumulative, N (%)	Percent, N (%)	Percent, N (%)	Cumulative, N (%)	Percent, N (%)	Percent, N (%)	Cumulative, N (%)	Percent, N (%)	Percent, N (%)	Cumulative, N (%)	Percent, N (%)	Percent, N (%)	Cumulative, N (%)	Cumulative, N (%)
±5							0.76 (3)	100 (394)	0.76 (3)	0.76 (3)	100 (394)	1.27 (5)	1.27 (5)	100 (394)	100 (394)

“ ” each short form in this table (i.e. FIM_8SF, FIM_4SF, MDS_13, MDS_8SF, MDS_4SF)

Weighted Kappa, Kappa, McNemar’s test and ICC between FIM_13 and the varied test forms (MDS_13, MDS_8, MDS_4)

Table 4

Test Form	p value	Weighted Kappa statistics	Agreement strength	ICC	ICC strength
Stroke (<i>n</i> = 1065)					
MDS_13	<0.0001 *	0.73	Strong	0.90	Excellent
MDS_8	<0.0001 *	0.73	Strong	0.91	Excellent
MDS_4	<0.0001 *	0.69	Strong	0.88	Excellent
Lower limb amputation (<i>n</i> = 472)					
MDS_13	0.04 *	0.53	Moderate	0.70	Good
MDS_8	0.01 *	0.54	Moderate	0.70	Good
MDS_4	0.01 *	0.48	Moderate	0.65	Good
Knee replacement (<i>n</i> = 568)					
MDS_13	0.0001 *	0.17	Weak **	0.40	Fair **
MDS_8	<0.0001 *	0.14	Weak **	0.35	Poor **
MDS_4	0.0016 *	0.09	Weak **	0.22	Poor **
Hip replacement (<i>n</i> = 394)					
MDS_13	<0.0001 *	0.55	Moderate	0.80	Excellent
MDS_8	<0.0001 *	0.44	Moderate	0.76	Excellent
MDS_4	<0.0001 *	0.34	Fair	0.67	Good

p value^: p value from McNemar’s Test for amputation FRG due to 2 * 2 table computation

* Significant agreement between FRGa and FRGc

** Biased result due to insufficient variability of the data

Table 5

The coefficient of variation of method error (CVME) for each functional related group (FRG) diagnosis

Test form	CV _{ME}
Stroke ($n = 1065$)	
MDS_13	20.92
MDS_8	21.00
MDS_4	22.63
Lower limb amputation ($n = 472$)	
MDS_13	4.46
MDS_8	4.48
MDS_4	4.71
Knee replacement ($n = 568$)	
MDS_13	2.47
MDS_8	2.69
MDS_4	4.13
Hip replacement ($n = 394$)	
MDS_13	10.00
MDS_8	11.37
MDS_4	12.90