View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Brunel University Research Archive

Rasch Analysis of the Upper-Limb Sub-scale of the STREAM Tool in an Acute Stroke Population

Dr. Bilal A. Mateen MBBS

University of Warwick, Warwick Medical School, Coventry, UK

Coventry, CV4 7AL, United Kingdom

Email: b.mateen@warwick.ac.uk

Telephone: +44 (0)20 7679 2000

Dr. Karen Baker PhD

University of Hertfordshire, Department of Health and Social Work, Hertfordshire, UK

Hertfordshire, AL10 9AB, United Kingdom

Email: kyacobi@gmail.com

Telephone: +44 (0)1707 284000

Prof. E. Diane Playford* MD FRCP

University of Warwick, Warwick Medical School, Coventry, UK

Coventry, CV4 7AL, United Kingdom

Email: D.playford@warwick.ac.uk

Telephone: +44 (0)2476 573273

Author for correspondence (*)

Manuscript Word Count: 3,983

Abstract Word Count: 184

Acknowledgement & Declarations –The authors have no conflict of interests to declare.

Funding - This study was funded by the Stroke Association (grant number TSA 2007/14).

Ethics Statement - Ethical approval was obtained from the University College London Institute of Neurology Joint Research Ethics Committee and the National Hospital for Neurology and Neurosurgery

Abstract

Background – Stroke is a leading cause of disability worldwide. The most common impairment resulting from stroke is upper limb weakness.

Objectives - To determine the usefulness and psychometric validity of the upper limb sub-scale of the STREAM in an acute stroke population.

Methods: Rasch Analysis, including unidimensionality assumption testing, determining model fit, and analysis of: reliability, residual correlations, & differential item functioning.

Results - 125 individuals were assessed using the upper limb sub-scale of the Stroke Rehabilitation Assessment of Movement (STREAM) tool. Rasch analysis suggests the STREAM is a unidimensional measure. However, when scored using the originally proposed method (0-2), or using the response pattern (0-5) neither variant fit the Rasch model ($p < 0.05$). Although, the reliability was good (Person-Separation Index – 0.847 & 0.903 respectively). Correcting for the disordered thresholds, and thereby producing the new scoring pattern, led to substantial improvement in the overall fit (chi-square probability of fit - 22%), however, the reliability was slightly reduced (PSI – 0.806).

Conclusions - The study proposes a new scoring method for the upper limb sub-scale of the STREAM outcome measure in the acute stroke population.

Word Count: 184

Key Words: Psychometrics, Stroke, Patient Outcome Assessment , Upper Extremity, Neurological Rehabilitation

Introduction

Stroke is a leading cause of disability worldwide.[1] Despite improvements in acute medical care following stroke, more than 250,000 people live with disabilities caused by stroke.[2] The most common impairment resulting from stroke is upper limb weakness,[3] which can impact self-care, work, and leisure activities. Therefore, upper limb rehabilitation plays an important role in improving long term outcome.[4]

A problem commonly encountered by clinicians is the selection of the most appropriate outcome measure to assess physical impairment due to stroke, and improvement as a result of rehabilitation.[7] This is because there is a vast array of potential tools available, and it can be difficult to discern between their clinical utility without formally assessing their validity. The two most commonly utilized approaches for outcome measure validation are Classical Test Theory (CTT), and Rasch Model Theory (RMT).

The Stroke Rehabilitation Assessment of Movement (STREAM) is a tool used to assess rehabilitation outcome in stroke patients (see box 1 for a description of the STREAM tool).[5] Several studies have illustrated that the STREAM is both reliable and valid based on a CTT approach,[5,8-9] however this validation method has recently come under criticism for its theoretical and practical limitations.[10-11]

Hsueh and colleagues performed a Rasch analysis on all of the subscales of the STREAM in a chronic (median time – 12.5 months post-event) stroke population.[12] This produced a smaller 15 item STREAM-S measure. However, the upper limb subscale of the STREAM has not been analyzed to determine its psychometric properties in an acute/sub-acute stroke population. One continuing source of discussion in the literature is the importance of timing with regards to rehabilitative interventions, i.e. do some interventions produce greater improvements in motor function if conducted during the acute phase, rather than the chronic.[13] To be able to effectively answer this question the tools used to measure

change, such as the STREAM, must be psychometrically robust in both populations.

The purpose of this study was to provide a Rasch-model based analysis of the upper limb sub score of the STREAM outcome measure to determine its usefulness and validity in measuring upper limb function for acute stroke patients undertaking rehabilitation. Additionally, we sought to determine the optimal scoring method by comparing the two different methods of scoring the STREAM: 1) the original 3-point ordinal scale which disregards the qualitative 'abc' distinctions proposed by Daly et al.[6]; and 2) the 5-point ordinal scale which includes the 'abc' distinction.

Box 1 - The Stroke Rehabilitation Assessment of Movement (STREAM) Tool

The STREAM consists of three 10 item sub-scales: the upper limb, lower limb and basic mobility scales. Each item in the three sub-scales is scored using an ordinal scale. The limb scales are scored on a 3-point ordinal scale (0, 1a/b/c, 2), however, the final scoring system does not account for the a,b,c criteria attached to the score of 1, i.e. each is awarded a score of 1 regardless of the letter score. The inclusion of the a, b, & c criteria alongside the score of 1, was made as a qualitative distinction and included due to rater confusion that was identified during validation. Thus, the total possible score is 20 for each of the limb sub-scales, which can subsequently be transformed to a score out of 100 to correct for missing items that occur due to pain or a limited range of motion. The items of the mobility sub-scale are scored on a 4-point ordinal scale, with a maximum possible score of 40 points. During the development of the tools it was demonstrated that the three sub-scales can be used individually or in-combination;[6] for the purposes of this study we used only the upper-limb portion as the intervention delivered by the recruiting services was upper-limb specific. The upper limb portion of the Stroke Rehabilitation Assessment of Movement (STREAM) focuses on voluntary movement which utilize different muscle groups in and around the upper extremity, for example, protraction of the scapula.

Methods

Location

The three unique locations included in this study were: the Hyper-acute Stroke Unit (HASU – acute in-patients only), the Albany Rehabilitation Unit (ARU –acute and chronic in-patients), and the Neuro-Rehabilitation Unit (NRU –acute and chronic in-patients) at the National Hospital for Neurology and Neurosurgery (NHNN – a UK tertiary neurological centre).

Participants

An observational cohort was established using sequential recruitment of patients admitted to the aforementioned locations between July 2009 and July 2011.

Inclusion/Exclusion Criteria

Patients were 18+ years of age, had an imaging-confirmed diagnosis of stroke, and were less than 12 weeks post-stroke at the time of assessment for inclusion in this study. Gross screening of the participants for suitability for suitability for inclusion in the study was conducted by a research nurse. The participants were informally assessed to determine whether they had sufficient cognitive ability and language/communication skills to follow the instructions required to complete the STREAM. Each patient provided full informed to participate. Patients unable to read or with difficulties understanding the instructions (due to severe cognitive or language/communication impairment) were excluded. Pain, and multiple strokes were not considered exclusion criteria for this study, however bilateral pathology was.

Assessment

The STREAM has been described in detail in Box 1. A second upper extremity specific outcome measure, The Chedoke ARM and Hand Activity Inventory (CAHAI) was utilized to aid in characterizing the range of impairments in the sample under investigation. The CAHAI is a valid and reliable measure with 13 tasks (opening a jar of coffee, calling 911, pouring a glass of water, etc.), scored on a 7-point ordinal scale ranging from the individual requiring total assistance to complete the task (0), to compete independence in task completion (7).[16]

Patients admitted to any of the aforementioned locations completed the STREAM and CAHAI as part of a routine battery of admission outcome measures, regardless of presence or extent of upper limb dysfunction. Manual preference was confirmed by the participant during the assessment. The STREAM and CAHAI were administered and scored (in English) by an experienced and appropriately trained clinician (author - K.B.), who provided instructions and support during completion. Rehabilitative interventions were subsequently delivered by a team of qualified physiotherapists during the course of the participant's admission. Descriptive statistical analysis of the data was conducted using SPSS.[17]

Rasch Analysis *[Conducted using the unrestricted (partial credit) model in RUMM 2030.[18] The parametrization of the item estimates in RUMM2030 is described elsewhere.[19]]*

Rasch analysis is a post-dictive method of psychometric analysis, which can be thought of as a probability-based analysis that determines the degree to which a pattern of observed responses corresponds to/fits the pattern predicted by the Rasch model.[14] Rasch analysis is often used to assess the structure and measurement properties of outcome measurement tools, specifically those that produce categorical data such as the STREAM tool. Assuming that specific criteria are fulfilled, the process of Rasch analysis identifies the relative difficulty of each item in a tool, and separately determines each individual's relative

skill/impairment with regards to what the tool is aiming to measure. Given that the Rasch model assumes the probability of selecting or affirming a specific score on an item of a questionnaire depends on the patient's degree of impairment/skill, and the inherent difficulty of that action/task, it is therefore possible to ascertain whether the outcome measure in question performs as the model predicts. And subsequently, post-hoc corrections to the tool can be made to improve fit to the Rasch model. More in-depth discussions pertaining to the underlying mathematical model or the process of Rasch analysis can be found in the following citations.[14-15]

Fit Statistics

The primary statistic used to evaluate how well an outcome measure fits the Rasch model is the χ^2 item-trait interaction statistic. This value represents the sum of the χ^2 values for each item in the scale. The probability of fit is derived on the basis of the sum total of the degrees of freedom. Acceptable fit is described as a non-significant χ^2 probability value, which for this study was set at the 5% level ($p = 0.05$).[20] The secondary statistic used to assess how well the items fit the Rasch model are the item fit residual statistics. Statistical evaluation of this statistics is based on the residual values, where misfit is illustrated by fit residual values of more than ±2.5 and/or χ^2 p value below the Bonferroni adjustment significance threshold. The Bonferroni adjustment is a conversion applied to the significance threshold value (e.g. $p = 0.05$) to reflect the number of items being considered. Each individual Bonferonni adjustment is stated with the results, and the base probability value utilized to calculate the alpha is always $p = 0.05$. The summary fit residuals for the items and persons are included for the for the original and the final re-scored version of the STREAM.

Threshold order

The transition point between each score (i.e. 0 to 1a, 1a to 1b, etc.) are known as thresholds, and they reflect the point at which there is equal probability of an individual being classified into two adjacent categories.[21] Within the STREAM there are 5 potential categories for each question, and therefore 4 thresholds. The purpose of Rasch analysis is to identify where the categories and thresholds perform in a manner predicted by the model. Where there is a discrepancy between the observed response pattern and predicted pattern, the threshold appears disordered, and thus the probability of a specific score is never high enough for there to be a transition point. To correct this problem different response categories in an item can be collapsed to produce a single new category, and the outcome of this change can be monitored using the fit statistics to determine whether the change was beneficial.

Reliability

Two different reliability parameters have been calculated. The first statistic is the person separation index (PSI), which indicates the degree of reliability of the fit statistics.[21] Moreover, it illustrates the STREAM's ability to discriminate between individuals with different levels of upper limb weakness/impairment. A result in excess of 0.7 is deemed sufficient to be able to differentiate across at least three patient groups.[22] The second reliability statistic is the Cronbach's α. Whilst the latter statistics is more commonly used in CTT psychometric analysis, it requires case-wise deletion of individuals with missing values, and thus reduces the amount of information available in the sample. The minimum acceptable α value is 0.7.[23]

Test of Unidimensionality

The unidimensionality assumption is one that refers to the presumption that a single factor is being measured. As such, if an outcome measure is unidimensional, then it should be

possible to theoretically place all of the items in order of difficult with regards to that single factor. Unidimensionality was tested using the method originally described by Smith.[24] A 95% confidence interval was then generated using a binomial test to define the proportion of tests that fail to meet the criteria of unidimensionality. A result consistent with a unidimensional scale will have the lower bound of the 95% confidence interval as less than or equal to 0.05.

Residual Correlations

A potential source of misfit is the presence of local dependency, where an individual's response on one item has some bearing upon their response to another item. Whilst there is no consensus in the literature concerning a specific value at which the correlation is significant, a common approach is that a residual correlation of 0.2 more than the average of all the item residual correlations can be considered problematic.

Item Characteristic Curves (ICC) and Differential Item Functioning (DIF)

Item Characteristic Curves are visual illustrations of the concordance between the observed scores for different ability levels (marked a points on a graph), and a curve representing the expected sores for a specific item. The relationship between the expected and observed values can be used to identify whether an item is prone to over-, or under-discrimination. Moreover, these curves can be used to determine whether there are underlying differences in response pattern based on additional variables, such as demographics (e.g. Sex), which is known as DIF. ANOVA tests were utilized to assess DIF, and a threshold of $p = 0.05$ was used to determine significance.

Comparing the Response Pattern to the Original Scoring Method

The above analysis was conducted on the dataset where the original 3-point ordinal scale scoring method proposed by Daly et al.[6], which disregards the qualitative 'abc' distinctions, was utilized. To determine if the 5-point ordinal scale which includes the 'abc' distinction is superior to the original, the data was reformatted so that instead of transforming the recorded scores from 0,1a,1b,1c,2 to 0,1,2, it became 0,1,2,3,4. All of the above Rasch analysis methods were then repeated on the new dataset.

Reporting Standards

This manuscript conforms to the STROBE reporting guidelines.

Results

125 patients who suffered a stroke of varying sub-types (table 1) were recruited to the study. Mean time from stroke to assessment was 3 weeks (S.D. 3 weeks). The mean age of the participants was 62.7 years (standard deviation – 17.7). The demographics of the study population are summarized in table 1. For the response frequencies see Table S1 in the supplementary material.

[Table 1]

Does the STREAM Questionnaire response pattern fit the Rasch Model?

The items of the STREAM Questionnaire were found to have a substantial degree of deviation from the Rasch model (Table 2 – Original). The item fit residual was -1.57 (S.D. – 1.69), and the associated chi squared test probability was <0.001. On closer examination (Table 3), 4 items (4, 5, 6, & 7) had residual fit values outside of the acceptable range (±2.5). Moreover, two items (1 & 3) had chi-squared probability values that were statistically significant suggesting they are extremely misfitting. Furthermore, all 10 items had disordering thresholds. In summary, the response pattern for the original version of the STREAM questionnaire does not appear to fit the assumptions of the Rasch model.

[Table 2 & 3]

Test of Unidimensionality

The 95% confidence interval for the proportion of tests that fail to meet the unidimensionality criteria is [0.036,0.113] suggesting that the upper-limb scale of the STREAM is a unidimensional scale.

Differential Item Functioning (DIF)

Bonferonni adjustment of the base probability (p = 0.05), where n = 30, resulted in a significance threshold of 0.001667. Analysis of the 10 items for uniform and non-uniform DIF by age and sex, illustrated there was no significant variation by either demographic variable with regards to the response pattern on the outcome measure.

Residual Correlations

The 10 items demonstrated a high degree of redundancy, illustrated by the several statistically significant levels of correlation between the questions (see Table S2 in the supplementary material). Only two items (3 & 7) did not have significant residual correlation with at least one other item in the scale.

The Original Scoring System for the STREAM Questionnaire

A set of summary statistics for the behavior of the original scoring pattern (0-2) upon analysis using the Rasch model is available in Table 2. The results clearly demonstrate that the original pattern demonstrates significant misfit with regards to the Rasch model. The probability values are both 0 at the number of decimal places reported by the RUM2030 program, for the two variants analyzed. Given the results thus far, we thought it was appropriate to consider re-scoring the STREAM from the original 3-point scoring pattern into a 5-point response pattern, inclusive of the 'abc' distinctions. This was done in an attempt to determine whether the STREAM in any format would fit the Rasch model.

Re-scoring the STREAM Questionnaire

The STREAM was re-scored (Table 4) to correct the disordered thresholds. The questions can be split into two groups based on the new response pattern: Items 1, 3, 4, 5, & 6 were changed from 01234 to 00112; and items 2, 7, 8, 9, & 10 were changed from 01234 to 01112. It should be noted that the 01112 response pattern is not actually different in terms of the score assigned to the individual in the original scoring pattern of the STREAM, as the a,b,c criteria do not translate into different scores; each is still only assigned a value of 1. The purpose of collapsing the response criteria is because the model identified that they were not (probabilistically) discriminative (see figure 1 for rationale). Each alteration was added in an iterative process to monitor the change in overall fit. Once the disordered thresholds were corrected, the overall fit to the Rasch model improved substantially (Table 2). The item fit residual degrades to -3.97 (S.D. 1.8388) from -1.57 (S.D. – 1.69), but the associated chi squared test probability of fitting the Rasch model improved to 0.222. On closer examination, no single item had a chi-squared probability value that was statistically significant, unlike previously (raw data not included). The patterns described above for the residual correlations, Item Characteristic Curves and DIF, whilst altered were not significantly different than the patterns described for the original version of the STREAM measure (raw data not included). Although, the reliability of the scale did decrease slightly, to a person-separation index of 0.81, whilst the Cronbach's α remained largely unchanged (Table 2).

[Table 4]

Discussion

The results of the study found that the upper limb sub-scale of the STREAM outcome measure in its original form, whilst being a unidimensional measure, did not fit the Rasch model. However, modifying the scoring system resulted in substantially better overall fit to the Rasch model, and was associated with good reliability indices (high Person-Separation Index and Cronbach's α). The analysis identified no differential item functioning, but did demonstrate substantial residual correlations between the 10 items in scale.The residual correlation results are unsurprising given that actions/movements will never completely isolate muscles, and thus, where co-operative action of these muscles occurs, the results will inevitably show high/significant correlations.

Moreover, the results demonstrated that the original developers of the STREAM outcome measure were correct in ignoring the a,b,c, criteria for most of the questions (items 2, & 7-10), as the re-scoring resulted in a similar pattern of scoring as the one described in the original development study (01112).[6] However, the descriptive thresholds at which 0 and 1 point were awarded in the original study do not appear to be consistent throughout the tool. Our analysis demonstrated that items 1, 3, 4, & 5 did not abide by the original scoring pattern. Instead, the optimal solution was that 1a was scored as 0, where it was previously assigned a score of 1 by the developers.[6] Interestingly, there is a notable clinical difference in the two clusters: items 2,7,8,9, & 10 are all movements that occurs from the elbow distally (i.e. flexion at the elbow, opening and closing the hand, etc.), whereas the other items all utilize the muscles of the back and shoulder (i.e. raising the arms overhead, shrugging shoulders, etc.). The modified rating criteria based on the results is described in Table 4.

Comparison to literature

Hseuh and colleagues, when they Rasch analyzed the STREAM in a chronic stroke

population, found that two items of the upper limb scale (1 and 3) did not fit the Rasch model, and therefore removed them from subsequent analysis.[12] Our initial results were similar (Table 3). However, upon re-scoring the items, the previously significant χ^2 values which suggested extreme misfit, no longer met the Bonferroni adjusted significance threshold (0.005000). Items 1 and 3's χ^2 values improved to 0.155650 and 0.006610, respectively. It would be interesting to determine whether a similar effect would have been observed if Hsueh and colleagues had chosen to re-score the scale, before excluding the items, as this information does not appear to have been reported.[12] Another key difference between the two studies, is the use of, and validation, of a single sub-scale in isolation of the other sub-scales. This methodological difference may explain some of the discrepancies between the findings described in this study and those described by Hsueh.[12] Future research should examine the behavior of all three sub-scales of the STREAM outcome measure being used simultaneously in an acute stroke population.

Strengths and Weakness

A potential weakness of the scale itself is that it appears to be unable to discriminate very well between individuals. Of the 125 individuals, 91 achieved the same score, which is represents 73.8% of the sample clustering at one point. An effect that is visible in the response pattern, and the Person-Item Map (see Figure S1 in the supplementary material). One potential reason for this is that the scoring system even after Rasch modification still produced a 3-point ordinal scale, where 0 signified absence or near absence of coordinated movement, and 2 was completely unimpaired movement. As such, any impairment that did not satisfy either of those extremes, which was most of the instances recorded, received the same score of 1, explaining the observed clustering.

The main potential weakness of this study is that the sample was drawn from a

tertiary center, which may limit the generalizability of the results. However, the activity and participation measure utilized as part of the admission battery, which is based on the degree of upper limb paralysis (CAHAI), suggested that the degree of disability tended towards the milder end of the spectrum (Table 1 – Demographics). 45% of the participants in this study achieved the maximum possible score on the tool, and the vast majority had scores in the upper half of the score range (0 to 91). This suggests that the degree of disability ranged from mild to moderate for most of the patients, and thus, the results are more widely applicable than the tertiary nature of the participating center would initially suggest. Alternatively, it is possible that the milder residual deficit was a result of the patients in this study being younger than the average stroke patient in the UK,[25-26] which could be another manifestation of the specialist nature of the recruiting center. Furthermore, we have assessed the results of the rasch analysis using the same data that we used to generate the results, which means that our observations could be the consequence of over-fitting. Genuine out-of-sample validity would require the use of a new dataset to test our results, which is an outstanding task currently.

Sample Size Calculation

The number of individuals required to establish stable person and items estimates using the Rasch model, is based on the degree of error expected. An analysis of sample sizes found that to achieve an item calibration stability of +0.5 logits with a 95% confidence interval is 100 individuals, and with a 99% confidence interval is 150 individuals.[20] As such, the sample size utilized in this study (n = 125), whilst it may appear relatively small, is more than adequate to drawn reasonable conclusions from.

Implications for Clinicians

Rasch analysis has allowed us to identify the interval scale that underlies the STREAM through a logarithmic transformation. A recent study using the Rasch analyzed version of the Fugl-Meyer Assessment demonstrated that these interval scale results can be used to accurately map standardized assessment results to appropriate short and long term rehabilitation goals.[27,28] As such, we believe that the continuous linear (interval) scale we have identified is likely to be much more useful to clinicians and policy makers than the ordinal values currently produced by the STREAM, as it more accurately reflects the relative difficulty of attaining each additional point on the scale. For example, the original ordinal scale would have you believe that the difference between an improvement from 0 to 3, and 12 to 15 is equal. However, the interval scale (see appendix) demonstrates this is not true. The true improvement from 0 to 3 is equal to 4.26 intervals, whereas from 12 to 15 is 2.18 intervals, almost half.

Conclusion

In conclusion this study proposes a new scoring method for the upper limb sub-scale of the STREAM outcome measure in the acute stroke population, which after correction for misfit to the Rasch occurring in its original form, resulted in a unidimensional, and highly reliable measure, which satisfied the expectations and assumptions of the Rasch model. However, the results illustrate quite substantial clustering of scores, which suggests that the clinical usefulness of this tool may be limited.

References

[1] – Adamson J, Beswick A, Ebrahim S. Is stroke the most common cause of disability?. Journal of Stroke and Cerebrovascular Diseases. 2004 Aug 31;13(4):171-7.

[2] – Carroll K, Murad S, Eliahoo J, Majeed A. Stroke incidence and risk factors in a population-based prospective cohort study. Health Statistics Quarterly. 2001;12:18-26.

[3] – Lawrence ES, Coshall C, Dundas R, Stewart J, Rudd AG, Howard R, Wolfe CD. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. Stroke. 2001 Jun 1;32(6):1279-84.

[4] – Han C, Wang Q, Meng PP, Qi MZ. Effects of intensity of arm training on hemiplegic upper extremity motor recovery in stroke patients: a randomized controlled trial. Clinical Rehabilitation. 2013 Jan;27(1):75-81.

[5] – Daley K, Mayo N, Wood-Dauphinee S. Reliability of scores on the Stroke Rehabilitation Assessment of Movement (STREAM) measure. Physical therapy. 1999 Jan 1;79(1):8.

[6] – Daly K. The STroke REhabilitation Assessment of Movement (STREAM): Content Validity and Preliminary Reliability. McGill University.1994.

[7] – Barak S, Duncan PW. Issues in selecting outcome measures to assess functional recovery after stroke. NeuroRx. 2006 Oct 31;3(4):505-24.

[8] – Hsueh IP, Hsu MJ, Sheu CF, Lee S, Hsieh CL, Lin JH. Psychometric comparisons of 2 versions of the Fugl-Meyer Motor Scale and 2 versions of the Stroke Rehabilitation Assessment of Movement. Neurorehabilitation and neural repair. 2008 Nov;22(6):737-44.

[9] – Hsueh IP, Wang CH, Sheu CF, Hsieh CL. Comparison of psychometric properties of three mobility measures for patients with stroke. Stroke. 2003 Jul 1;34(7):1741-5.

[10] – Baker K, Barrett L, Playford ED, Aspden T, Riazi A, Hobart J. Measuring arm function early after stroke: is the DASH good enough?. Journal of Neurology, Neurosurgery

126 & Psychiatry. 2015 Jul 15:jnnp-2015.
127 [11] – Mateen BA, Doogan C, Hayward K, Hourihan S, Hurford J, Playford ED. Systematic
128 review of health-related work outcome measures and quality criteria-based evaluations of
129 their psychometric properties. Archives of Physical Medicine and Rehabilitation. 2017 Mar
130 31;98(3):534-60.
131 [12] – Hseuh I, Wang WC, Wang CH, Shen CF, Lo SK, Lin JH, Hsieh CL. A simplified
132 stroke rehabilitation assessment of movement instrument. Physical therapy. 2006 Jul
133 1;86(7):936-43.
134 [13] – Takeuchi N, Izumi SI. Rehabilitation with poststroke motor recovery: a review with a
135 focus on neural plasticity. Stroke research and treatment. 2013 Apr 30;2013.
136 [14] – Andrich D. Rasch models for measurement. Sage; 1988.
137 [15] – Ehlan A, Kucukdeveci A, Tennant A. The Rasch Measurement Model. In: Franco
138 Franchignoni (Ed). Research Issues in Physical & Rehabilitation Medicine. Pavia: Maugeri
139 Foundation, 2010:89–102
140 [16] – Barreca S, Stratford P, Lambert C, Masters L & Streiner D. Test-retest reliability,
141 validity, and sensitivity of the Chedoke Arm and Hand Activity Inventory: A new measure of
142 upper-limb function for survivors of stroke. *Arch Phys Med Rehabil*. 2005;86: 1616-1622.
143 [17] - IBM Corp. Released 2013. IBM SPSS Statistics for Macintosh, Version 22.0. Armonk,
144 NY: IBM Corp.
145 [18] –Andrich D, Lyne A, Sheridan B, Luo G. RUMM 2030. Perth: RUMM Laboratory.
146 2012.
147 [19] – Andrich D, Luo G. Conditional pairwise estimation in the Rasch model for ordered
148 response categories using principal components. Journal of applied measurement. 2002
149 Dec;4(3):205-21.

[20] – Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Archives of physical medicine and rehabilitation. 1994 Feb;75(2):127-32.

[21] – 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 Care & Research. 2007 Dec 15;57(8):1358-62.

[22] – Fisher WP. Reliability statistics. *Rasch measurement transactions*. 1992;*6*(3):238.

[23] – Nunnally JC. Assessment of Reliability. In: Psychometric Theory (2nd ed.). New York: McGraw-Hill. 1978.

[24] – Smith Jr EV. Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. Journal of applied measurement. 2002.

[25] – Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of rehabilitation medicine. 2003 Jan 1;35(3):105-15.

[26] – Ferro JM, Crespo M. Young adult stroke: neuropsychological dysfunction and recovery. Stroke. 1988 Aug 1;19(8):982-6.

[27] – Playford ED, Siegert R, Levack W, Freeman J. Areas of consensus and controversy about goal setting in rehabilitation: a conference report. Clinical Rehabilitation. 2009 Apr 1;23(4):334-44.

[28] - Velozo CA, Woodbury ML. Translating measurement findings into rehabilitation practice: an example using Fugl-Meyer Assessment-Upper Extremity with patients following stroke. Journal of Rehabilitation Research & Development. 2011 Dec 15;48(10).

Table 1 – Demographics of Sample Population

Sex (n = 125)	
Male (n)	59.2% (74)
Female (n)	40.8% (51)
Age (n = 125)	
- 49	12.8% (16)
50-59	19.2% (24)
60-69	43.2% (54)
70 - 79	23.2% (29)
80 -	0.2% (2)
Handedness (n = 125)	
Right	96.0% (120)
Left	4.0% (5)
Stroke Location (n = 123, insufficient location information = 2)	
Right	**(62)**
ACA	5
MCA	25
PCA	6
Lacunar	22
Brainstem	4
Left	**(61)**
ACA	3
MCA	19
PCA	10
Lacunar	23
Brainstem	6
Arm function (CAHAI)	
Score on Outcome Measure	Number of Individuals
- 19	14
20 - 29	8
30 - 39	4
40 - 49	1
50 - 59	7
60 - 69	19
70 - 79	12
80 - 90	8
91	56

ACA – Anterior Cerebral Artery, MCA – Middle Cerebral Artery, PCA – Posterior Cerebral Artery

Table 2 – Summary Statistics for the STREAM and the Re-scored Variant of the STREAM

STREAM Original [Response Pattern Scoring]				
Person Separation Index	With Extremes (n = 125)	0.90		
	Without Extremes (n = 123)	0.84		
Item-Trait Interactions	Chi Square	71.156		
	Probability	<0.001 (Degrees of Freedom. – 20)		
Cronbach's Alpha	With Extremes (n = 111)	0.92		
	Without Extremes (n = 109)	0.90		
(Not including extremes)	***Items***		***Persons***	
	Location	**Fit Residual**	**Location**	**Fit Residual**
Mean	0.00	-1.57	0.39	-2.35
Standard Deviation	1.17	1.69	1.17	2.20
STREAM Original [Original Scoring]				
Person Separation Index	With Extremes (n = 125)	0.85		
	Without Extremes (n = 123)	0.76		
Item-Trait Interactions	Chi Square	96.41		
	Probability	<0.001 (Degrees of Freedom. – 20)		
Cronbach's Alpha	With Extremes (n = 111)	0.91		
	Without Extremes (n = 109)	0.89		
(Not including extremes)	***Items***		***Persons***	
	Location	**Fit Residual**	**Location**	**Fit Residual**
Mean	0.00	-2.78	0.79	-2.44
Standard Deviation	1.99	3.12	2.17	1.69
STREAM Re-scored [Rasch-based Novel Scoring System]				
Person Separation Index	With Extremes (n = 125)	0.81		
	Without Extremes (n = 121)	0.73		
Item-Trait Interactions	Chi Square	24.489		
	Probability	0.222 (Degrees of Freedom. – 20)		
Cronbach's Alpha	With Extremes (n = 111)	0.92		
	Without Extremes (n = 108)	0.88		
(Not including extremes)	***Items***		***Persons***	
	Location	**Fit Residual**	**Location**	**Fit Residual**
Mean	0.00	-3.96	0.08	-3.01
Standard Deviation	0.43	1.84	1.44	2.13

Table 3 – Logit Location, Fit Statistics, and ICC description of Individual STREAM Items (Pre-Correction of Disordered Thresholds)

Item	*Task*	*Location*	*SE*	*Fit Residual*	*Chi Square*	*Degrees of Freedom*	*Chi Square p Value*	*ICC*
1	**Supine** *Protracts scapula in supine*	-2.24	0.17	1.04	18.65	2	<0.001*	5
2	**Supine** *Extends elbow in supine*	0.49	0.15	-1.12	2.54	2	0.282	1
3	**Sitting** *Shrugs shoulder (Scapular elevation)*	-2.17	0.18	1.39	23.65	2	<0.001*	5
4	**Sitting** *Raises hand to touch top of head*	0.40	0.13	-3.00	0.89	2	0.642	2
5	**Sitting** *Places hand on sacrum*	0.45	0.13	-3.42	1.51	2	0.470	2
6	**Sitting** *Raises arm overhead to fullest elevation*	0.50	0.12	-3.19	1.44	2	0.488	2
7	**Sitting** *Supinates and pronates forearm*	0.32	0.12	-2.52	5.01	2	0.082	2
8	**Sitting** *Closes hand from fully opened position*	0.66	0.09	-2.37	3.29	2	0.193	1
9	**Sitting** *Opens hand from fully closed position*	0.88	0.09	-1.52	5.64	2	0.060	1
10	**Sitting** *Opposes thumb to index finger*	0.73	0.08	-1.03	8.54	2	0.014	1

SE – Standard Error. ICC – Item Characteristic Curves: 1 - Marginal over-discrimination; 2 – Classic over-discrimination; 3 – Classic fit; 4 – No systemic deviation, but individual class intervals deviate from the model; 5 – Marginal under-discrimination; and 6 – Classic under-discrimination.
** Probabilities below the Bonferroni adjusted p value (adjusted value = 0.001 for 10 items from probability base of 0.01)*

Table 4 – Scoring Patterns for the Re-scored Variant of the STREAM

Item	Original Response Pattern	New Scoring Pattern	Corresponding Descriptions for New Scoring Pattern
1	0/1a/1b/1c/2	0/0/1/1/2	**<u>0/0/1/1/2</u>** 0 – Unable to appropriately perform the test movement (includes completing part of the movement but with marked deviation in ability compared to the unimpaired side). 1 – Patient completes part of the action in a manner similar to the unimpaired side OR completes the entire action but with marked deviation in ability compared to the unimpaired side. 2 – Patient completes action in a manner similar to the unimpaired side.
2		0/1/1/1/2	
3		0/0/1/1/2	
4		0/0/1/1/2	
5		0/0/1/1/2	
6		0/0/1/1/2	**<u>0/1/1/1/2</u>** 0 – Unable to perform the test movement, or any part of it. 1 – Patient is capable of completing part of, or the entire test movement, but with marked deviation in ability compared to the unimpaired side. 2 – Patient completes action in a manner similar to the unimpaired side.
7		0/1/1/1/2	
8		0/1/1/1/2	
9		0/1/1/1/2	
10		0/1/1/1/2	

Figure Legends

Figure 1: Probability Curves for Items 1, 4 and 8

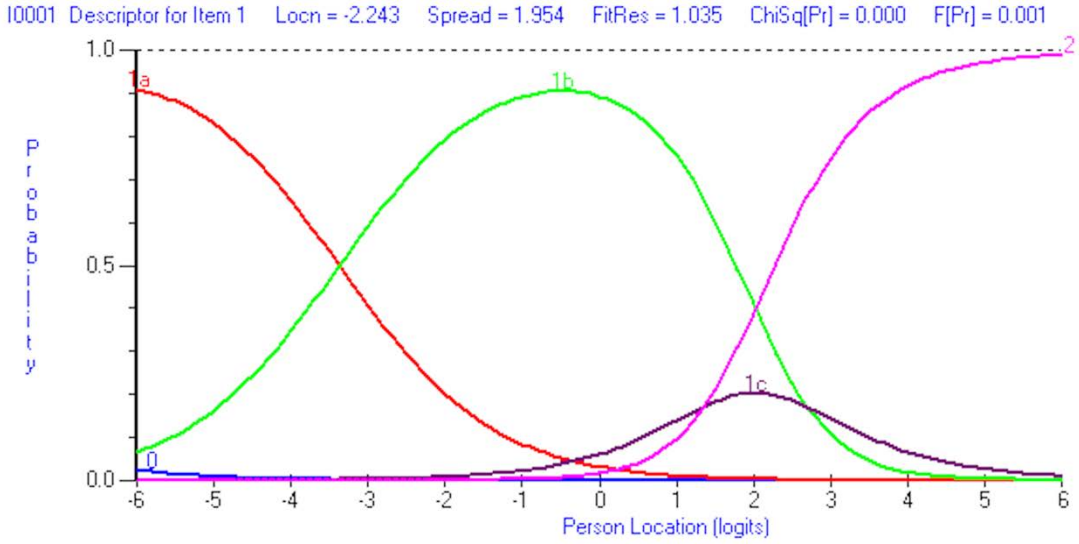
Fig. 1 A and B correspond to item 1 before and after re-scoring, respectively. Before re-scoring, options 1a and 1c are the cause of the disordered threshold. Using the corresponding descriptions for these scores (1a = “able to perform only part of the movement, and with marked deviation from unaffected pattern”, 1b = “able to perform only part of the movement, but in a manner that is comparable to unaffected side”, 1c = “able to complete the movement, but only with marked deviation from unaffected pattern” ,[6] the optimal solution identified was to combine these 3 options resulting in the scoring pattern 01112. This suggests that for these questions, the distinction between only completing part of the movement, and the full movement (assuming impairment is noted), is not sufficiently different in terms of difficulty for the measure to discern, and thus re-scoring was necessary.

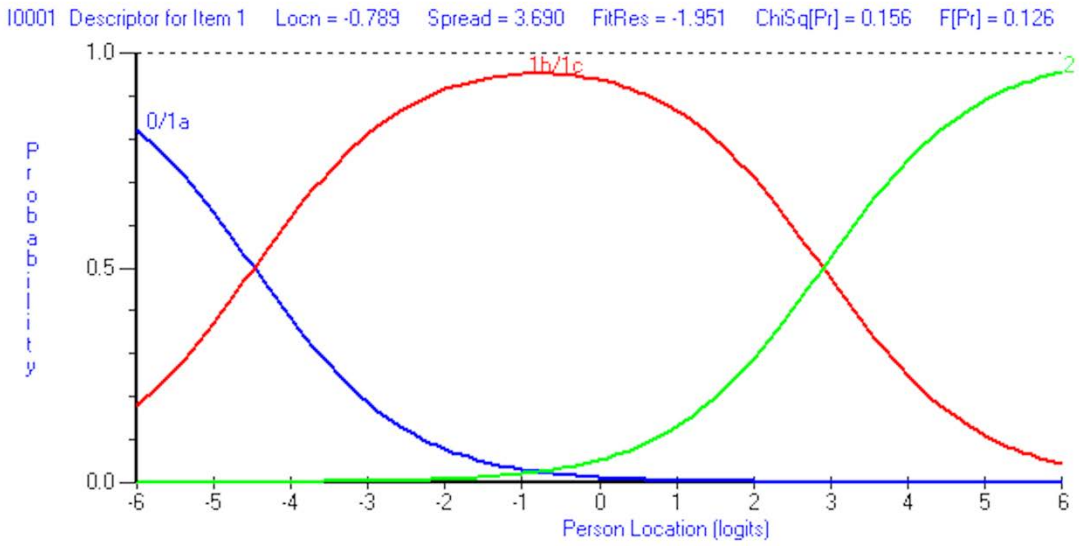
Fig. 1 C and D correspond to item 4 before and after re-scoring, respectively. Before re-scoring, options 1a and 1c are the cause of the disordered threshold. Using the corresponding descriptions for these scores (see above), the optimal solution identified was to combine these 3 options resulting in the scoring pattern 01112. This suggests that for these questions, the distinction between only completing part of the movement, and the full movement (assuming impairment is noted), is not sufficiently different in terms of difficulty for the measure to discern, and thus re-scoring is necessary.

Finally, Fig. 1 E and F correspond to item 8 before and after re-scoring, respectively). Before re-scoring, only option 1a appears to be the cause of the disordered threshold. However, the optimal solution identified was to combine 1a, 1b and 1c, resulting in the scoring pattern 01112, instead of just combining 1a and 1b. This suggests that for these questions, the distinction between the ability to perform part of the action (1b), and the complete action (1c)

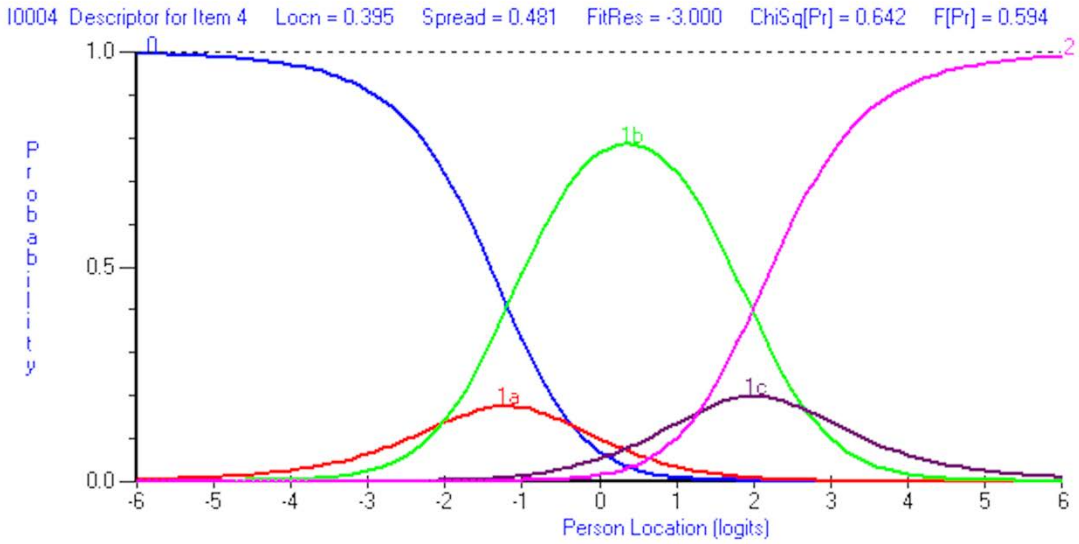
is sufficiently different, however, discounting this additional information meant that the item fit the Rasch model better overall.

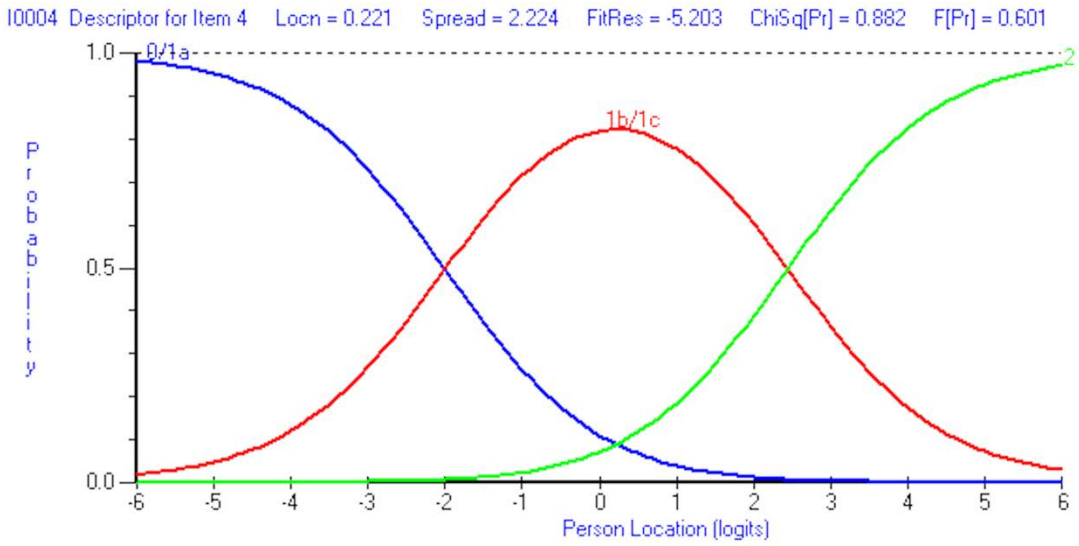
Figures 1 A &B


I0001 Descriptor for Item 1 Locn = -2.243 Spread = 1.954 FitRes = 1.035 ChiSq[Pr] = 0.000 F[Pr] = 0.001
Probability
1.0
0.5
0.0
1a
1b
1c
0
2
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Person Location (logits)

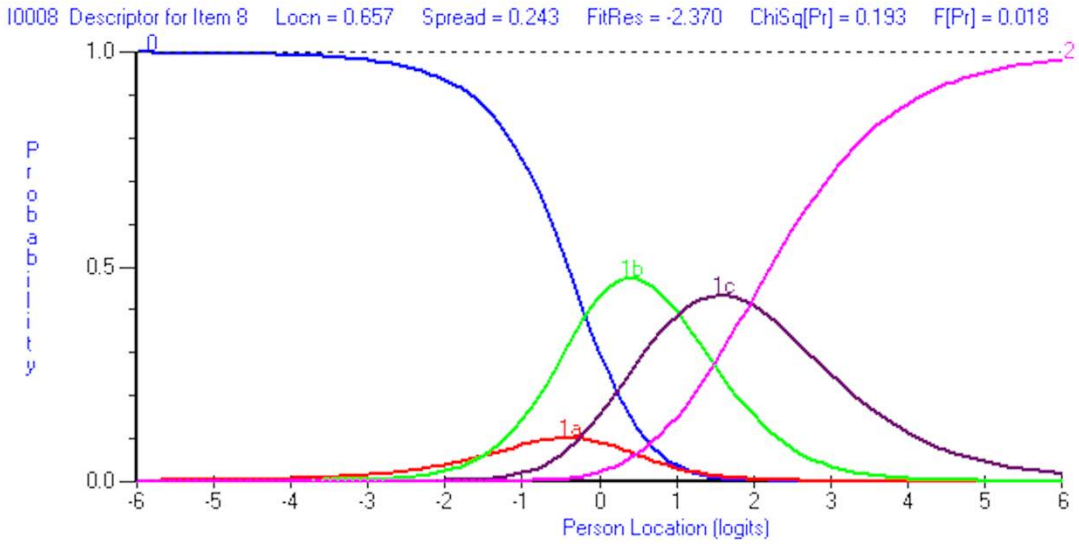

I0001 Descriptor for Item 1 Locn = -0.789 Spread = 3.690 FitRes = -1.951 ChiSq[Pr] = 0.156 F[Pr] = 0.126
Probability
1.0
0.5
0.0
0/1a
1b/1c
2
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Person Location (logits)

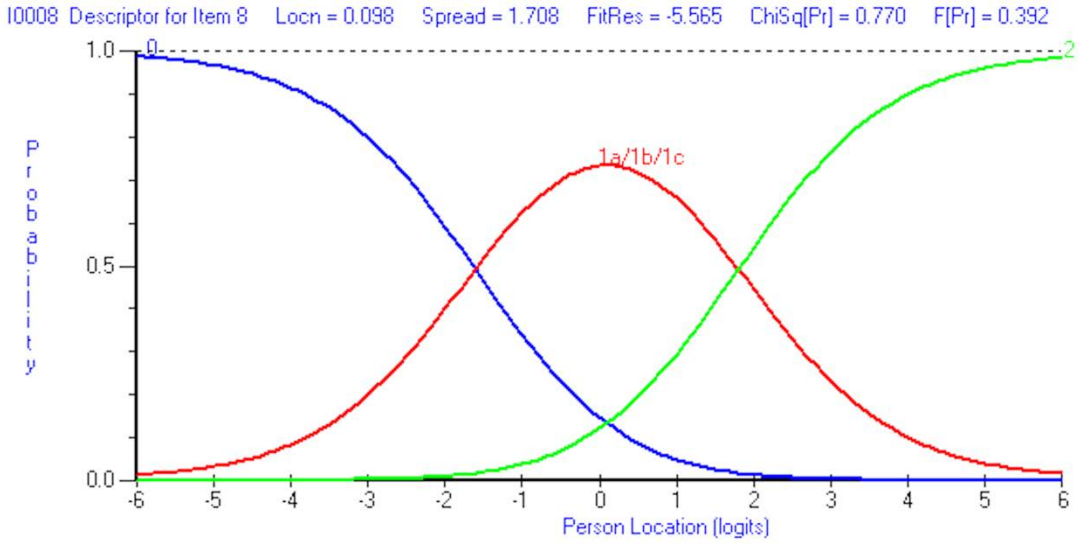
Figures 1 C & D


I0004 Descriptor for Item 4 Locn = 0.395 Spread = 0.481 FitRes = -3.000 ChiSq[Pr] = 0.642 F[Pr] = 0.594
Probability
1.0
0.5
0.0
0
1a
1b
1c
2
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Person Location (logits)


I0004 Descriptor for Item 4 Locn = 0.221 Spread = 2.224 FitRes = -5.203 ChiSq[Pr] = 0.882 F[Pr] = 0.601
Probability
1.0
0.5
0.0
0/1a
1b/1c
2
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Person Location (logits)

Figures 1 E & F


I0008 Descriptor for Item 8 Locn = 0.657 Spread = 0.243 FitRes = -2.370 ChiSq[Pr] = 0.193 F[Pr] = 0.018
Probability
1.0
0.5
0.0
0
2
1a
1b
1c
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Person Location (logits)


I0008 Descriptor for Item 8 Locn = 0.098 Spread = 1.708 FitRes = -5.565 ChiSq[Pr] = 0.770 F[Pr] = 0.392
Probability
1.0
0.5
0.0
0
2
1a/1b/1c
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Person Location (logits)