Is it time to revisit the Boston Carpal Tunnel Questionnaire? New insights from a Rasch model analysis

Christina Jerosch-Herold, MSc, PhD[1]

Jeremy DP Bland, MD, ChB[2]

Mike Horton, PhD[3]

[1]Faculty of Medicine and Health Sciences, University of East Anglia, Norwich Research Park, Norwich, UK

[2]Dept of Neurophysiology, East Kent Hospitals University NHS Foundation Trust, Canterbury, Kent, UK

[3]Psychometric Laboratory for Health Sciences, Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, UK.

Acknowledgements: none

Keywords: Item response theory, psychometrics , Boston Carpal Tunnel Questionnaire, outcome measures, Rasch Model

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mus.27173

Corresponding author: Professor Christina Jerosch-Herold, School of Health Sciences, University of East Anglia, Norwich Research Park, Norwich, UK, Tel: 01603593316, Email: c.jerosch-herold@uea.ac.uk Twitter: @tinajerosch

Ethical Publication Statement: We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

Disclosure of Conflicts of Interest: None of the authors has any conflict of interest to disclose.

Funding statement: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Abstract

Introduction: The Boston Carpal Tunnel Questionnaire (BCTQ) is a patient-reported outcome measure (PROM) used to measure symptom severity and function in carpal tunnel syndrome

(CTS). Despite its wide usage, investigation of its measurement properties using modern psychometric methodologies is limited.

Methods: Completed BCTQ data collected routinely in the Canterbury carpal tunnel clinic was used to investigate the structural validity and measurement properties of the BCTQ through application of a Rasch model analytic approach.

Results: A total of 600 patients with electrodiagnostically confirmed CTS in their right hand were randomly selected from the database and analysed. Mean age was 48.8 years and 73% were women. Initial analysis showed that the 19 items could not be reliably added up to form a single linear construct. All subsequent analyses were done by subscale only.

The Symptom Severity subscale (SSS) displayed a large amount of local dependence. This could be accommodated through the creation of four clinically-derived testlets, allowing for the ordinal SSS raw score to be transformed to a linear measure.

The Function subscale displayed a number of issues regarding its psychometric integrity. These include scale and item fit, targeting, differential item functioning, and dimensionality.

Discussion: This study shows that a single total score generated across all BCTQ items is not psychometrically valid, and that the Symptom and Function subscales should be treated separately. We propose a modified scoring system for the Symptom subscale, resulting in a linear measure that can be used in the analysis of future and existing datasets.

INTRODUCTION

Outcome assessment of carpal tunnel syndrome interventions often focuses on patient-reported symptom resolution and improvement in function [1]. One of the most widely used patient-reported outcome measures (PROM) is the Boston Carpal Tunnel Questionnaire [2]. It comprises two subscales: a symptom severity scale and a functional status scale which together assess symptoms during the day and night, and difficulty with everyday tasks, respectively. Each item is scored on a five-point ordinal scale, where a higher score represents more severe symptoms or worse function. A final score is derived from each subscale as the average of the 11 or eight items, respectively ranging from 1.0 to 5.0 [3]

The BCTQ is easy to administer, free and has been translated into several languages. It has been extensively tested for concurrent validity and test-retest reliability using classical test theory [4]. Exploratory and confirmatory factor analysis confirm a two-factor solution consistent with the two subscales for symptoms and function [5]. However, others have found that a three-factor model was a better fit [6]. A shorter version called the CTS-6 was developed using item response theory [7,8] in which the 11 items from the symptom severity scale are reduced to six items. It is the first study to apply modern psychometric methods, namely Rasch model analysis, a form of item-response theory suggested to improve measurement accuracy in ordinal level scales [9].

Rasch model analysis offers an alternative, and arguably much richer, description of performance of PROMs at both item and scale level compared to classical test theory [10]. The premise of the Rasch model is that the answer chosen by a subject to a question about the severity of a symptom will be determined both by the severity of that symptom in that individual (termed 'person ability' or person's level of a trait) and by how well the available answers represent the possible range of symptom severity (termed 'item difficulty'), and that both characteristics can be measured on the same 'ruler'. The process essentially tests whether it is valid to sum a set of items into a total score, for a given population. If a set of items suitably fit the Rasch model, then it is also possible to convert their total summed score (which is always ordinal) into an interval-level measurement that can be legitimately used in parametric statistical procedures (e.g. to determine change scores or effect sizes, etc.) [11]. This is particularly important in clinical trials where BCTQ scores are

used as the primary outcome to test treatment efficacy. Rasch model analysis is useful in evaluating existing scales to assess how appropriate it is to use total scores as outcome measures, and whether ordinal raw scores can be converted to continuous data. Additionally, it is especially useful for developing new scales, and for constructing item banks for computer-adaptive testing (CAT) such as those used by PROMIS, which allows fewer questions to be used without losing test precision. [12]

To date only one study has applied Rasch model analysis to the BCTQ [7]. Their Rasch analysis did not explore response thresholds or differential item functioning. They proposed a shorter version of the symptom severity scale including only six items, and did not explore alternative ways by which the full 11-item scale can be converted into true interval level measurement. The latter is particularly useful in analysing existing datasets based on the full 11-item version, which continues to be widely used.

More recently an alternative version of the full BCTQ was proposed based on decision tree modelling, called DT-BCTQ [13] which reduced the BCTQ to three items in both the symptom severity scale and functional status scale. One major disadvantage of their approach is that the scoring algorithm relies on computer administration, which may not always be feasible in a clinical setting.

The objective of this secondary analysis was to evaluate the structural validity of the full BCTQ and its two subscales using Rasch model analysis in order to identify strengths and weaknesses and whether modifications are needed.

MATERIALS AND METHODS

Data source

Data from 600 patients were randomly selected from from the East Kent CTS database which contains 30,000 patient records. Patients in this database had been asked to complete the BCTQ separately for the right and left hands at every clinic attendance.

The diagnosis of CTS was based on clinical history, signs and symptoms and verified by nerve conduction studies. NCS severity was classified according to Bland criteria[14] . Ethics approval for the use of anonymized data extracted from the database for studies of this type was obtained from the London-South East Committee of the NHS National Research Ethics Committee.

Minimum sample sizes of at least 150 are recommended for Rasch analysis, whilst a ratio of at least 10 respondents per item, or 10 respondents per category for polytomous items, are suggested for item calibration to be reliably stable [15]. Our sample size of 600 is more than adequate to determine the relative 'difficulty' ordering of all items and thresholds, especially when the two domains of the BCTQ are considered separately.

Inclusion criteria: We selected the right hand in right-handed patients with either unilateral right-sided CTS or bilateral CTS, which was either more severe in the right hand or symmetrical between the hands, where the neurophysiological severity of the right-side CTS was grade 1-3 [14]. Age was dichotomized as ≤62 years or 63 years and above and recorded with sex and CTS grade. Patients who had completed the BCTQ on more than one occasion were only included once and incomplete questionnaires were excluded.

Rasch analysis

Rasch analytic procedures are described in detail elsewhere[16 17]. The complete process that we followed is provided in Supplementary file I, which broadly follows the methodological processes as described by

others [11 18]. Briefly, this process determined whether all items in the scale essentially measure the same construct (unidimensionality); whether the items maintain a consistent 'difficulty' ordering across the full range of the underlying construct (item fit); how well-aligned the scale and population are, in terms of the construct being measured (targeting); whether items are related too closely to each other, above what is expected from the underlying construct (local dependence); whether the item response categories are working correctly (response category threshold ordering); whether the items can reliably place the people being measured into the correct order (reliability); and whether any of the items behave differently between specific independent groups (e.g. by age group, or by sex) (differential item functioning – DIF). A glossary of terms with explanations is provided in Appendix I.

All Rasch analyses were performed using RUMM2030 software (Rumm Laboratory; http://www.rummlab.com.au) for windows 10.

RESULTS

Complete BCTQ data from 600 patients were available. Their mean age was 48.8 (SD = 12.8) years and 162 (27%) were men. Nerve conduction severity distribution was: 30.8% were very slight (grade 1), 34.5% were mild (grade 2) and 34.7% were moderate (grade 3)[14]. Symptom severity score group median was 2.73 (IQR: 2.27; 3.27) and mean was 2.77 (SD = 0.72). For the Functional Severity score group median was 1.99 (IQR: 1.25; 2.5) and mean was 1.99 (SD= 0.84).

The initial Rasch analysis of all 19 items of the BCTQ together showed significant misfit with the Rasch model as well as a lack of unidimensionality (Table 1, analysis A1).

A principal components analysis (PCA) of the residuals should show only random patterns in how the items cluster (load) together. However, in this case the loading patterns showed distinct separation of the symptom and functional scales, with only Q11 of the symptom scale (Do you have difficulty with grasping and use of small objects such as keys or pens?) loading alongside the functional items.

This indicated that it is not appropriate to sum these subscales together to form a single total score, as they represent different constructs. We therefore proceeded to analyse each subscale separately.

Symptom Severity subscale (SSS)

1. *Distribution of responses*

All five responses categories were used in all 11 items. Use of the lowest response category exceeded 15% (+3% allowance) [19] in three items. Q11 (grasping small objects) showed the strongest ceiling effect (39% of the sample endorsing 'no difficulty at all'). The proportion of respondents using the worst category (5) was very low for the items on numbness and weakness, however this is not surprising given that the sample was limited to those with mild to moderate disease severity only.

2. *Response thresholds*:

Five items displayed disordered thresholds: Q1, Q2, Q4, Q5 and Q10. Two analytic approaches were taken at this point: an approach which focussed on the functionality of the response options, and the correction

of this through response category collapsing (described in full in supplementary file I); and an approach which focuses on correcting for local dependence first, which offers functional benefits in terms of clinical pragmatism (reported in the main manuscript).

3. Local dependence.

We found local dependence in 10 pairs of items, with several well in excess of 0.1 (mean correlation = -0.1).

4. Unidimensionality

The SSS did not meet the assumption of a unidimensional scale as 19.4% of the series of *t*-tests were significant at 5% level (lower bound confidence interval = 17.7%).

As local dependence can contribute towards apparent multidimensionality, we inspected the dependency-grouping of the items to look for clustering patterns. This led to the formulation of four super items or 'testlets', where the grouping of items into testlets was based on a combination of statistical evidence (local dependence pairs and factor loading) and subjective clinical judgement regarding subdomains. The four testlets were composed as follows: night symptoms and nocturnal waking (Q1, Q2, Q9 and Q10), daytime pain severity, duration and frequency (Q3, Q4 and Q5), numbness and tingling (Q6 and 8) and weakness and dexterity (Q7 and Q11).

A subsequent series of t-tests procedure found only 5.18% of t-tests significant at $p<0.05$ (95%CI: 3.4-6.9%) indicating a unidimensional scale (Table 1, analysis 3).

5. Item fit

The modified 4-testlet SSS showed overall good fit with the Rasch model (Chi-square statistic= 47.52 (df=36), p=0.09. Individual Item fit residuals for the four testlets were all within ±2.5. Response thresholds were found to be disordered for testlets 1 (night symptoms and nocturnal waking) and 2 (daytime pain severity, duration and frequency). However, no additional rescoring should be conducted on a testlet, as total scores on a single testlet item can be formed in different ways and there is no formal response

structure, whereas for individual items a respondent can only select one of a number of given response options.

6. Differential item functioning

There was no evidence of significant uniform or non-uniform DIF by sex (male; female) or age group (≥ 63 years; 64 plus), for any of the 11 individual items or the four testlets.

7. Reliability

The person separation index after creating testlets is 0.72 indicating that the scale can distinguish between at least two subgroups. Although the complete 11-item set has a higher person separation index of 0.86, this is artificially inflated by the dependency that is present.

8. Targeting

A person-item threshold map demonstrates a good match between the distribution of items on the SSS with person's severity, indicating good targeting of the SSS to this sample (Supplementary Figure 1).

9. Linear transformation

When the assumptions of the Rasch model are satisfied, the sufficiency of the raw score allows for the transformation to a linear measure. This transformation table is presented (Supplementary File II) where the original raw ordinal scale score is converted into an interval logit score, along with a corresponding raw-score equivalent that corresponds to the original range. Please note that this conversion is conditional on complete data (no missing responses).

<u>Functional Status Subscale (FSS)</u>

1. Distribution of responses

The FSS shows strong ceiling effects with all 8 items having no difficulty endorsed by at least 25% of respondents.

2. Response thresholds:

Response thresholds were ordered for all 8 items indicating that the five response categories worked well and respondents can distinguish between these.

3. Local dependence

There was some evidence of dependence in three pairs of items. The largest dependency was between items 3 'holding a book' and 4 'gripping a telephone handle', with a further dependency between items 1 'writing' and 2 'buttoning of clothes'. There was a further apparent dependency between items 2 and 8 'bathing and dressing'. All of these dependencies seem logical in terms of their content.

4. Unidimensionality

An equating t-test between two subsets of positively and negatively loaded items was just outside the acceptable threshold for a unidimensional scale (7.1% [95%CI: 5.3 to 9.0%]) (see Table 1, Analysis 4), although this is a reflection of the dependency that is present. A conventional Principal Component Analysis (PCA) was also carried out, which supports a stable 1-factor solution where the first component explained 65% of variance.

5. Item fit

Rasch fit statistics for individual items were non-significant and residuals were within ±2.5 logits, with the exception of 2 items: Q6: household chores (fit residual -2.7, p=0.005) and Q7: carrying shopping (Chi-Square statistic p=0.001). The overall item fit statistic for the FSS was statistically significant (p<0.001) indicating misfit.

6. Differential item functioning

Statistically significant uniform differential item functioning was found by age on item 2 'doing up buttons' where older people were more likely to report problems, and item 4 'gripping a telephone handle' where

younger people were more likely to report problems. Uniform DIF by sex was also detected for item 5 'opening jars', where women were more likely to report problems.

7. Reliability

The Person Separation Index is 0.85 indicating the ability of the FSS to distinguish between at least 3 subgroups, although this is likely to be artificially inflated by the dependency that is present.

8. Targeting

The FSS does not seem to be very well targeted to the persons as can be seen by the large gap in items relative to the persons' ability which is skewed to the left of the item-threshold distribution diagram (Supplementary Figure 2).

DISCUSSION

This Rasch model analysis provides insights into the construct validity of the BCTQ, which has implications for what can be inferred from the total score of the BCTQ and its two subscales. Firstly, summing all 19 items into a single score violates the assumptions of the Rasch model. Instead the two subscales should be reported as separate scores.

The Symptom Severity Scale did not entirely meet the expectations of a linear structure as proposed by the Rasch model. The reasons were manifold: firstly, response thresholds for some items especially those reporting the frequency of nocturnal awaking from pain and tingling were disordered. This could indicate that patients cannot distinguish between 5 different frequencies of waking at night. Although this could be resolved to some extent through a post-hoc rescoring process to reduce five original response categories to three functioning response categories, the scale also displayed high item-to-item residual correlations indicating local dependence. Several items ask about the same symptoms, for example pain severity, pain duration and pain frequency. The presence of local dependence could itself be a cause of disordered thresholds as well as inflating the reliability (person separation index). By creating 'testlets', where items are combined into a 'super-item', dependence can be accommodated without the need to rescore or remove items. Dependence can also drive multidimensionality [20], and the creation of four testlets for the symptom subscale resulted in a unidimensional scale. The grouping of items into testlets was based on clinical reasoning alongside the statistical evidence from the Rasch model analysis.

The SSS did not show any response bias by age or sex, and reliability as indicated by a high PSI and targeting of item difficulty to person ability is satisfactory.

When deleting the same items as Atroshi et al [7] did and combining the two questions on daytime tingling and numbness time into a testlet we still found significant misfit to the Rasch model, including disordered thresholds and further local dependence. A further advantage of retaining all 11 items of the SSS and accommodating dependence by creating testlets is that the usual method for calculating the total score can be used (adding 11 item responses and dividing by 11) without having to rescore individual items. This is particularly important in a busy clinical setting.

Finally it allows a raw score to logit (interval) transformation (provided in Supplementary file II), which means it is a true interval scale and parametric statistics can be used when undertaking efficacy analysis in clinical trials.

The Functional Status Scale showed overall misfit to the Rasch model. Its strengths are that all response thresholds were ordered, and it is borderline unidimensional. The unidimensionality finding concurs with Atroshi et al. [8] who also found that the FSS fitted a one-factor structure thus measuring a single latent trait of function. However, we did observe some strong ceiling effects with the categories indicating greater difficulty being used very little, although this is to be expected given that the sample only included patients with NCS severity grades 1 to 3, where motor function remains largely unaffected. This is also reflected in the targeting of person ability to item difficulty within this sample, which is misaligned. The results also show response bias by age for some items which means that scores may need to be disaggregated by sex and age when calculating group means. It is also possible that the FSS measures other disabilities which are not CTS related, for example in the item on bathing and dressing. Others [8] have suggested that the FSS may not be specific enough to CTS and other validated region-specific PROMS which measure function should be used (e.g. Quick Disabilities of Arm, Shoulder and Hand [21] or the Michigan Hand Questionnaire [22]). It is over 25 years since the original BCTQ was conceived and items in the FSS such as 'gripping telephone handle' may be considered out of date. The use of smaller digital home and mobile phones places very different demands on the hand such as a dexterous and sensate thumb which may explain the apparent response bias by age group. As the FSS showed several sources of misfit to the Rasch model and due to the availability of newer psychometrically valid PROMs for assessing hand function we did not explore further solutions to achieve better fit.

Although we retained the original response categories and scoring across all items, there may still be issues with the response category structure of items 1, 2, 4, 5 and 10.

It should also be considered that all analyses were carried out using data from the English-language version of the scale, and these findings may not be consistent across alternative cultural and language-adapted formats. Additionally, it must be noted that any post-hoc amendments, as are carried out in analysis

software, are implied, and it may be that patients would respond differently when presented with items in a different order or with fewer items such as those used in the CTS-6.

Conclusions

Our secondary analysis of this large dataset of routinely collected BCTQ scores in CTS patients has shown that the SSS and FSS scale should not be summed into a total score but treated as two separate subscales. A revised Rasch version of the SSS which shows good fit to the Rasch measurement model has allowed us to generate a raw score to logit scale transformation. It is this transformed score which should be used when analysing efficacy data from trials.

The FSS does not fit the Rasch measurement model. The content validity of the FSS is questionable and items are not specific enough for CTS. We therefore propose the complementary use of alternative, psychometrically valid PROMS to assess hand function in CTS patients.

List of Abbreviations:

BCTQ Boston Carpal Tunnel Questionnaire
CTS carpal tunnel syndrome
DIF differential item functioning
FSS Functional Status Subscale
NCS nerve conduction studies
PCA principal component analysis
PROM patient-reported outcome measure
PSI person separation index
SSS Symptom Severity Subscale

REFERENCES

1. Jerosch-Herold C, Leite JCdC, Song F. A systematic review of outcomes assessed in randomized controlled trials of surgical interventions for carpal tunnel syndrome using the International Classification of Functioning, Disability and Health (ICF) as a reference tool. *BMC musculoskeletal disorders* 2006; **7**: 96.

2. Levine D, Simmons B, Koris M, et al. A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. *The Journal of Bone and Joint Surgery* 1993; **75A**(11): 1585-92.

3. Jongs R. Carpal Tunnel Questionnaire. *J Physiother* 2017; **63**(2): 119.

4. Mehta SP, Weinstock-Zlotnick G, Akland KL, Hanna MM, Workman KJ. Using Carpal Tunnel Questionnaire in clinical practice: A systematic review of its measurement properties. *Journal of hand therapy : official journal of the American Society of Hand Therapists* 2020.

5. Leite JCdC, Jerosch-Herold C, Song F. A systematic review of the psychometric properties of the Boston Carpal Tunnel Questionnaire. *BMC musculoskeletal disorders* 2006; **7**: 78.

6. Lue YJ, Wu YY, Liu YF, Lin GT, Lu YM. Confirmatory Factor Analysis of the Boston Carpal Tunnel Questionnaire. *J Occup Rehabil* 2015; **25**(4): 717-24.

7. Atroshi I, Lyren PE, Gummesson C. The 6-item CTS symptoms scale: a brief outcomes measure for carpal tunnel syndrome. *Qual Life Res* 2009; **18**(3): 347-58.

8. Atroshi I, Lyren PE, Ornstein E, Gummesson C. The Six-Item CTS Symptoms Scale and Palmar Pain Scale in Carpal Tunnel Syndrome. *J Hand Surg-Am* 2011; **36A**(5): 788-94.

9. Narayanaswami P, Burns TM. Clinical outcome assessments: The "Rasch-Ionale" for improved accuracy. *Muscle & nerve* 2018; **58**(3): 327-9.

10. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health technology assessment* 2009; **13**(12): iii, ix-x, 1-177.

11. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *The British journal of clinical psychology / the British Psychological Society* 2007; **46**(Pt 1): 1-18.

12. Jayakumar P, Overbeek CL, Vranceanu AM, et al. The use of computer adaptive tests in outcome assessments following upper limb trauma A SYSTEMATIC REVIEW. *Bone Joint J* 2018; **100b**(6): 693-702.

13. Jansen MC, Evers S, Slijper HP, et al. Predicting Clinical Outcome After Surgical Treatment in Patients With Carpal Tunnel Syndrome. *The Journal of hand surgery* 2018.

14. Bland JDP. A neurophysiological grading scale for carpal tunnel syndrome. *Muscle & nerve* 2000; **23**(8): 1280-3.

15. Linacre JM. Sample Size and Item Calibration [or Person Measure] Stability. *Rasch Measurement Transactions [Internet]* 1994; **7**(4).

16. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute of Educational Research; 1960.

17. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. *Journal of rehabilitation medicine* 2011; **43**(10): 884-91.

18. Yorke J, Horton M, Jones PW. A critique of Rasch analysis using the Dyspnoea-12 as an illustrative example. *J Adv Nurs* 2012; **68**(1): 191-8.

19. de Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: a practical guide. Cambridge, UK: Cambridge University Press; 2011.

20. Prodinger B, O'Connor RJ, Stucki G, Tennant A, Network II. Establishing Score Equivalence of the Functional Independence Measure Motor Scale and the Barthel Index, Utilizing the International Classification of Functioning, Disability and Health and Rasch Measurement Theory. *Journal of rehabilitation medicine* 2017; **49**(5): 416-22.

21. Kennedy CA, Beaton DE, Smith P, et al. Measurement properties of the QuickDASH (disabilities of the arm, shoulder and hand) outcome measure and cross-cultural adaptations of the QuickDASH: a systematic review. *Qual Life Res* 2013; **22**(9): 2509-47.

22. Chung KC, Pillsbury MS, Walters MR, Hayward RA. Reliability and validity testing of the Michigan Hand Outcomes Questionnaire. *J Hand Surg-Am* 1998; **23A**(4): 575-87.

Table 1: Overall scale fit at different analysis stages

	Stages of analysis	n=	Mean item fit residual mean (SD)	Mean person fit residual (SD)	Item-trait total chi-square		PSI	Test of unidimensionality[1] (95%CI)
					χ^2 (df)	*P*		
	Ideal values		***mean=0, SD<1.4***	***mean=0, SD<1.4***		***>0.05***	***>0.85***	***<5%***
A1	Initial analysis of all 19 items	600	-0.05 (3.2)	-0.34 (1.5)	513.6 (171)	<0.001	0.91	19.06% (17.3; 20.8)
A2	Symptom severity scale (SSS) only (11 items)	600	0.36 (1.4)	-0.37 (1.4)	188.2 (99)	<0.001	0.86	19.4% (17.7; 21.1)
A2a	Symptom severity scale only, post-rescore	600	0.08 (1.13)	-0.47 (1.54)	181.7 (99)	<0.001	0.85	16.89% (15.1; 18.6)
A3	SSS create 4 testlets[a]	600	0.67 (0.8)	-0.36 (1.1)	47.5 (36)	0.095	0.72[b]	5.18% (3.4; 6.9)
A3a	SSS create 4 testlets[a], post-rescore	600	0.25 (0.49)	-0.44 (1.13)	30.4 (36)	0.731	0.74[b]	6.52% (4.8; 8.3)
A4	Functional Status scale (FSS) only (8 items)	600	-0.27 (1.8)	-0.39 (1.2)	93.9 (48)	<0.001	0.85	7.1% (5.3; 9.0)

[a] Subtest 1: Q1, Q2, Q9 & Q10; subtest 2: Q3, Q4 & Q5; subtest 3: Q6 & Q8; subtest 4: Q7 & Q11)

[b] PSI after creating testlets

Legend: df – degrees of freedom, FSS Functional Status Scale, SD = standard deviation, SSS Symptom Severity Scale, PSI Person Separation Index,