BMJ Open

Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-035436
Article Type:	Original research
Date Submitted by the Author:	31-Oct-2019
Complete List of Authors:	Protopapa, Eva; London School of Hygiene and Tropical Medicine, Department of Health Services Research & Policy van der Meulen, Jan; London School of Hygiene and Tropical Medicine, Department of Health Services Research & Policy Moore, Caroline; University College London, Division of Surgery and Interventional Sciences Smith, Sarah; London School of Hygiene & Tropical Medicine, HSRP
Keywords:	UROLOGY, Prostate disease < UROLOGY, SURGERY

Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis

Evangelina Protopapa[1] Jan van der Meulen,[1] Caroline M Moore,[2] Sarah Smith[1]

1 London School of Hygiene and Tropical Medicine, London UK

2 University College London, London, UK

Author email addresses:

e.protopapa@ucl.ac.uk

JanvanderMeulen@lshtm.ac.uk

caroline.moore@ucl.ac.uk

Sarah.Smith@lshtm.ac.uk

Address for correspondence;

Dr Sarah Smith, Associate Professor in Psychology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine 15-17 Tavistock Place, London WC1H 9SH. Tel: 0207 9272038

Email: sarah.smith@lshtm.ac.uk

Word count:

Abstract: 296

Text: 4044

Keywords:
PROM, psychometric, Rasch, prostate, surgery

ABSTRACT

OBJECTIVES: To use Rasch analysis to evaluate the psychometric properties (and identify specific anomalies to be resolved) of the urinary and sexual functions scales of the STAR instrument for use in clinical practice with individual men.

DESIGN: Prospective cohort study

SETTING: 9 UK surgery centres in secondary care

PARTICIPANTS: 403 men diagnosed with prostate cancer and completed at least one questionnaire immediately before and at 1 or 3 months after a radical prostatectomy.

INTERVENTIONS: Radical prostatectomy.

PRIMARY AND SECONDARY OUTCOMES: STAR instrument before and 1 and 3 months after their surgery.

RESULTS: Both urinary (7 items) and sexual function (6 items) had disordered thresholds, suggesting that the response categories are not working as intended. In the urinary scale, 3 items and in the sexual function scale, 5 items showed problems with item fit. Both scales showed items that were unstable over time (DIF by time). The urinary function scale showed 1 pair of items and the sexual function scale had 5 pairs of items that had item response dependency. Overall, reliability was acceptable at the group level for both scales. However, targeting was poor for both scales, indicating an inadequate match between location of items and the distribution of the patients. This suggests that the underlying constructs that the scales purport to measure are not clear.

CONCLUSION: Using Rasch analysis as a diagnostic tool, we identified that both the urinary and the sexual function scales have issues that need to be resolved before STAR can be used with confidence in clinical practice. The sexual function scale in particular is unlikely to provide precise estimates for the outcomes experienced by men after radical prostatectomy. These results demonstrate the need to evaluate the suitability of any PROM before implementation in routine clinical practice, preferably using modern psychometric methods**.**

STRENGTHS AND LIMITATIONS OF THIS STUDY

- used state of the art psychometric methods to determine if it is appropriate to use a total function score to describe a patient's sexual or urinary function.
- determined how well the items in each score reflect the experience of men who report the questionnaire
- determined specific anomalies in the scores that suggest that the scales are not being used and understood in the way that was intended
- did not change the items in the questionnaire based on our findings and so did not evaluate any potential improvement such changes would make

INTRODUCTION

The use of patient-reported outcome measures (PROMs) has rapidly increased (1-3). In the UK, PROMs are routinely collected for several areas of elective surgery to evaluate the outcomes in *groups of patients*, receiving a particular treatment or treated in a specific hospital (4, 5). Similar approaches are under consideration for other conditions.

However, there is a lack of evidence about the extent to which clinicians can use PROMs to make their clinical practice more responsive to *individual patients'* needs. Also, it has been suggested that PROMs can play an important role for patients as they can help to inform ways in which patients can self-manage their condition (6, 7).

A web-based tool known as STAR ('Symptom Tracking and Reporting') (8) has been developed at the Memorial Sloan-Kettering Cancer Center (New York, US) to monitor outcomes of radical prostate cancer treatment in individual patients. This instrument is used to inform both surgeons and men about functional outcomes after surgery, such as urinary, sexual and bowel function improvement or deterioration. Its development is just one example of the implementation of PROMs in prostate cancer practice to inform both clinicians and patients (9-11).

The STAR instrument was not designed to compare men's functional status before and after surgery because different questions are included in the pre- and post-treatment STAR questionnaires. This means that the assessment before surgery is on a different 'ruler' compared to after surgery and therefore there is no clear way of understanding what the change means. However in practice, for example in the English national PROMs programme, pre- and post-treatment PROMs are often compared to monitor the impact of elective surgery (2).

Instruments such as STAR aim to measure specific 'constructs'. It is important these instruments have adequate psychometric properties, otherwise they may produce scores that are 'inaccurate' (prone to systematic error) or 'imprecise' (prone to random error), making it difficult to understand what the observed scores mean and even more difficult to interpret changes over time.

The criteria that must be met to ensure that PROMs are robust are well established (12-15). They ensure that the 'scale' that results from adding up responses to individual questions ('items') relates to a clear underlying construct, as distinct from descriptive responses or simple counts of how many times a symptom occurs.

Like most health-related PROMs, the STAR instrument has been developed using traditional psychometric methods based on classical test theory (CTT). There are important limitations to these methods (16). First, the scales developed using CTT produce 'ordinal scores', where the difference between two adjacent scores at different points on the scale may not be equal. This poses a problem because most statistical analyses assume scores have interval properties where differences between adjacent scores are equal across the entire scale. When scales are based on ordinal scores, changes over time are especially difficult to interpret. Second, the scores can only be interpreted for groups of patients, because measures of statistical uncertainty of these scores (e.g. 'standard errors') are only computed at group level, which limits their use for individual patients (17). Third, the performance of scales is dependent on the particular sample in which they are used. This makes it difficult to compare studies and, even more importantly, undermines further the interpretation of changes over time.

Modern psychometric methods, such as those based on ‘item-response theory’ (IRT) or ‘Rasch measurement theory’, provide a way of overcoming these challenges. Both are mathematical modelling approaches transforming ordinal scales into interval measures, provided that certain model-related criteria are met. But whereas IRT takes a statistical approach of adding parameters to the model in order to improve its fit to the data, the Rasch paradigm takes a theory-driven approach that investigates why the data do not fit the Rasch model (18-20). The Rasch paradigm, however, keeps central the conceptual underpinning of the instrument and provides a clear set of diagnostic statistics that can help to identify anomalies in its scores.

Instruments developed using these modern psychometric methods have four main advantages over CTT-based instruments. First, they have the potential to generate truly interval scores, thus improving the accuracy and precision with which change over time can be evaluated. Second, measures of statistical uncertainty can be estimated for scores of individual respondents, meaning that the interpretation of scores at patient level is more meaningful. Third, it is possible to produce scales that do not depend on a particular sample’s characteristics. Fourth, they can create a model that contains both pre-and post-surgery items, and therefore all items can be calibrated on the same ruler. The usual pre and post-treatment scores can still be derived but calibrated in such a way that they can be properly compared.

In a systematic review of seven prostate cancer-specific PROMs, including the STAR instrument (21), we identified that modern psychometric methods had not been used to evaluate the psychometric properties of these instruments. In this study, we therefore used Rasch analysis to estimate urinary and sexual function for individual men based on responses to the STAR instrument that were provided by men immediately before and up to three months after radical prostate cancer surgery. In so doing, we identified anomalies that should be addressed to make the STAR instrument, or any other PROM that aims to monitor changes in outcomes over time after prostate cancer surgery, suitable for use in routine clinical practice.

METHODS

Setting and participants

Participants were recruited between November 2015 and March 2017 from nine centres that perform radical prostatectomy by any method (open, laparoscopic-assisted or robotic-assisted) in the UK. Men were eligible if they were diagnosed with prostate cancer, scheduled to have a radical prostatectomy, and had sufficient English language to understand the information about the study and complete the required online questionnaire.

The clinical team at each centre identified and approached eligible patients, informed them about the study, and registered those who were interested in taking part on the secure online portal. Registered patients received their login details by text or email and logged on to the portal to complete the consent form. Once patients had consented, they were directed to the online questionnaire. Patients were invited to complete the questionnaire before surgery, and at one, three, six and 12 months after surgery.

Instrument

The STAR instrument consists of four domains: sexual function, urinary function, bowel function, and overall quality of life. Our analysis focused on the urinary and sexual function scales obtained immediately before and one and three months after radical prostate cancer surgery. We excluded the bowel scale from psychometric analyses as with only two items it had insufficient content to be considered a scale. Likewise, the single-item scale for overall quality of life was not considered in our analysis.

Urinary and sexual function items are scored on 3 to 11-point Likert scales. The pre-surgery form of the STAR instrument includes seven urinary function items and the post-surgery form includes five. Two of these are the same across both forms (questions 2 and 4). For sexual function, the same six items are included in both pre- and post-surgery forms. Item scores are summed for the urinary and sexual function domains to give total domain scores, which are transformed to scores ranging from 0 to 100.

We made two wording changes to the STAR instrument. First, our data collection also included the EPIC-26 questionnaire (not reported in the present paper) which overlaps with some STAR items. Where an item existed in both questionnaires, we used the EPIC wording. These minor wording changes are unlikely to substantially change the performance of the item. Second, the standard updated version of STAR has a time frame of six months pre-operatively for both sexual and urinary function, four weeks post operatively for sexual function and one week post operatively for urinary function. To ensure consistency across time for both urinary and sexual function domains, we used a 4-week recall period throughout. We considered this long enough for all problems to be noticed and/or resolved.

All items were administered at all time points, but analysis was conducted on the combinations of items proposed by the original STAR instrument as described above.

Rasch Measurement Theory

We performed analyses based on Rasch measurement theory to determine if it is appropriate to use a total function score to describe a patient's sexual or urinary function. As comparisons are often made between pre- and post-surgery scores, we aimed to determine if the seven pre-surgery and five post-surgery items could be placed on the same ruler. If they can then meaningful comparisons can be made across time. To do this, we 'stacked' the data, in other words, we added the baseline and follow-up scores for each patient as separate records (22).

The analyses aimed to answer a number of questions. First, has a measurement ruler been successfully constructed? Second, have the people been successfully measured? Third, is the scale-to-sample targeting adequate? The approach to each of these questions is explained briefly below. A more extensive explanation of Rasch measurement theory can be found in a number of recent overviews (23).

Has a measurement ruler been successfully constructed?

Item threshold ordering: Each of the items of the scale has multiple response categories which are scored to create a polytomous response (Likert scale). For a higher level of functioning, the probability of 'endorsing' a higher response category should increase and the probability of endorsing a lower response category decrease. If each response category in turn (0, 1, 2, 3, 4, 5) has the highest probability of endorsement with increasing levels of functioning, the 'thresholds' between the categories (0-1, 1-2, 2-3, 3-4, 4-5) show a logical order. Thresholds are the location on the scale where the two adjacent response categories have equal probability (50%) of endorsement.

Empirically, however, thresholds can be disordered (e.g. 0-1, 2-3, 1-2), indicating that the response categories do not work as intended. This can be because an item has ambiguous wording or has labels on the response scale that are not sufficiently distinct. We evaluated whether the response categories are working as intended by a visual inspection of the 'category probability curves'.

Item fit validity: The items of the scale must work together ('fit') as a conformable set both clinically and statistically. Clinically, the item ordering along the continuum should make sense and statistically the items need to satisfy specified criteria. Otherwise, it is inappropriate to sum item responses to reach a total score and consider the total score as a measure of the construct. When items do not work together ('misfit') in this way, the validity of a scale needs to be questioned.

We evaluated the fit of each item to the Rasch model by inspecting its 'fit residual' (acceptable range of +/- 2.5) and considering the related Chi-square value. We also assessed visually how closely the observed 'class interval mean scores' follow the expected values in the 'item characteristic curve'. Class intervals are groupings of approximately equal numbers of respondents who have about the same level of functioning.

Differential item functioning (DIF): Stability of the item locations is assessed by evaluating 'differential item functioning (DIF)'. DIF occurs when different groups within the sample, for example patients of different age, respond differently to an item, despite having the same level of functioning. Uniform DIF occurs when these differences are the same across the entire range of levels of functioning and is identified through an ANOVA main effect for 'person factors', for example age. Non-uniform DIF occurs when the differences are inconsistent across the range of level of functioning and is identified by an interaction between the person factor and the class intervals in ANOVA analyses.

In both the urinary and sexual function scales, we evaluated DIF by age, ethnicity, relationship status and number of co-morbidities. For items that were scored both before and after surgery (two items for the urinary function scale and all six items for the sexual functioning scale), we also evaluated DIF by time point.

Item-response dependency: The response to one item should not directly influence the response to another. If 'item response-dependency' happens, measurement estimates can be biased, and reliability, indicated by the 'person separation index', is artificially increased. Item-response dependency is evaluated by examining the residual correlations between the items after the Rasch factor they have in common has been partialled out. A correlation coefficient with a value larger than 0.30 indicates potential response dependency.

Have the people been successfully measured?

Reliability: Reliability was examined using the 'person separation index' which is a statistic comparable to the Cronbach's alpha, often used in traditional methods based on CTT. It quantifies how reliably the scale distinguishes between respondents. It is computed from the variation among person locations relative to the standard error of estimate for each individual respondent (16). Higher person separation index values indicate better reliability; a value >0.70 at group level and >0.85 at individual level indicates adequate reliability (20).

Is scale-to-sample targeting adequate?

'Scale-to-sample targeting' describes the match between the range of the construct measured by the items and the range of the construct in the sample of patients. This is evaluated by the 'person-item distribution' which compares the difference between 'person locations' and 'item threshold locations' on the underlying ruler, that captures for example urinary or sexual function. Any gaps in item threshold locations, in particular at the low and high ends of the scale, means that the functioning of respondents located in that gap area cannot be measured precisely. In other words, their scores will have a relatively large standard error of measurement, because their estimation is severely affected by missing information.

All p-values were adjusted for sample size (n=500) as Chi-square values are sensitive to sample size (24). Furthermore, Bonferroni corrections for multiple testing were also applied. All analyses were carried using RUMM 2030 (25).

RESULTS

Study sample

Overall, 971 men were eligible, of whom 873 were approached, 714 were interested and 431 men completed the online consent form, giving an overall recruitment rate of 44.4%.

Of the 431 patients who provided consent, 403 patients (93.5%) completed at least one valid questionnaire. A total of 366 valid questionnaires were completed at baseline, 222 questionnaires were completed at one month after surgery and 181 questionnaires at three months after surgery. Table 1 describes the characteristics of the 403 patients included in this analysis. These patients had a mean age of 63 years (SD 6.7; range 41 – 78 years), were predominantly white or white-British (79.7%), and were mostly married or living with a partner (76.7%).

Overall fit to the model

The overall Chi-square statistic indicated that neither the urinary function nor the sexual function scale fit the Rasch model (urinary function, p<0.001; sexual function, p<0.001).

Threshold ordering

Both urinary and sexual function scales had items with disordered thresholds, indicating that the response options were not working as intended. The urinary function scale had disordered thresholds for 7 of the 10 items. For these 7 disordered items, the category probability plots in Figures 2a-2g illustrate that this is mainly a problem with the middle response options, suggesting that the wording was not clear or that the difference between categories was not well understood. For example, for Q3 of the urinary function scale (‘Over the last 4 weeks, how often have you found you stopped and started again several times when you urinated?’) there is no point at which threshold 2 (‘About half the time’) and threshold 3 (‘Less than half the time’) are the most likely to occur. If the response options were working as intended, the probability of each threshold should come in order.

All six of the sexual function items are disordered. This means that none of the response scales are working as they were intended. Figures 3a-3f indicate that it is mainly thresholds 2 and 3 that are disordered, suggesting that the middle categories of the response scales are not well understood and may need to be re-worded.

Item fit

Both the urinary and sexual function scales contained items that did not fit the model, when considering together their fit residual, Chi-square value, and the item characteristic curve. One urinary function item (Q23) failed all three criteria (Table 2) indicating misfit to the model. Two further items failed one or two criteria (Q3 and Q7) indicating a broader problem with item fit.

Five sexual function items failed all three criteria (Table 2) and the remaining item failed one of the three criteria suggesting further problems with item fit.

Differential item functioning (DIF)

Overall, items in both scales were stable (invariant) across different groups for age, ethnicity, relationship status and number of co-morbidities. However, both scales contained items that were unstable across time, with the sexual function scale containing a greater number of unstable items.

One urinary item (Q23) showed uniform DIF across time points (p<0.001). Patients’ response to this item were systematically higher at 3 months post-op compared to 1 month post-surgery, despite having equal underlying levels of urinary function.

One sexual function item (Q9) showed uniform DIF by time (p<0.001), such that responses were systematically higher at baseline than the other time points. In addition, five sexual function items showed non-uniform DIF by time (Q9, Q10, Q11, Q12, Q13; all p<0.001).

Item-response dependency

Both scales contained pairs of items that were dependent on each other, but the sexual function scale showed greater local dependency. One pair of urinary function items showed local dependency: Q19 and Q21 (residual correlation = 0.32).

Four pairs of sexual function items showed local dependency with relatively high residual correlations: Q10 and Q11 (residual correlation = 0.30), Q12 and Q13 (residual correlation = 0.59), Q12 and Q14 (residual correlation = 0.55), Q13 and Q14 (residual correlation = 0.51).

Reliability

Reliability was acceptable at group level for both scales (urinary function scale: person separation index = 0.75; sexual function scale: person separation index = 0.82).

Scale-to-sample targeting

The person-item distribution of the urinary function scale was relatively poor (Figure 1a). Although the middle of the person distribution is reasonably well matched by items, both extremes of the distribution have few items. This means that for men located at the lower end of the scale (including many men at one month after surgery) and at the higher end of the scale (including many men before surgery) the level of functioning cannot be precisely measured.

The targeting for the sexual functioning scale was also poor (Figure 1b). In particular, most items are located in the centre of the scale whereas the distribution of people is quite wide. This means that the sexual function for men located at the higher end of the scale (often men before surgery) and the lower end of the scale (most of the men after surgery) is very imprecisely measured.

DISCUSSION

Our analyses have identified that neither the urinary function items nor the sexual function items from the STAR instrument can be placed on a common metric that is robust for comparisons before and after surgery. Furthermore, a number of anomalies have been identified that suggest the scales are not working as intended. There is an inadequate match between location of items and the distribution of the patients, suggesting that the underlying constructs that the scales purport to measure are not clear. Consequently, the items do not measure the men’s function very accurately. The response categories for many items are not consistently used, some items do not work with the others as a conformable set and some items are not stable over time.

These results indicate that in its current form the items in the STAR instrument do not provide an adequate ruler to monitor urinary or sexual function in clinical practice. These problems are likely to make the estimation of an individual patient’s outcome after surgery less accurate and precise and using the questionnaire in its current form therefore carries a risk of misrepresenting actual urinary and sexual outcomes.

Our results demonstrate that the risk of inaccurate estimation of outcomes using STAR is likely to be most pronounced for men with either very good or very poor outcomes. The poor scale-to-sample targeting, particularly for the sexual functioning scale, also means that this problem is exacerbated for men with better function before surgery and worse function after surgery, creating clear problems for the interpretation of change scores that are supposed to capture the impact of surgery. Further, both scales have items that showed DIF by time providing further evidence that it is not meaningful to compare scores before and after surgery or compare scores taken at different times after surgery.

In the short term, some of the identified deficiencies can be addressed using post-hoc statistical techniques to re-score the disordered thresholds (16, 20) or to resolve for the uniform DIF (23) and item-response dependency (20). However, a more robust solution would be to conduct qualitative research with men who have experienced radical prostatectomy to understand why the questions are not well understood and why the response options are not used in the way that was intended. Qualitative research should also explore which areas of content are missing and how items could be formulated to address these gaps. A revised version based on these findings would then need to be psychometrically evaluated again to determine how well the amendments to content and scoring have addressed the identified problems.

This study is the first to use robust modern psychometric methods such as Rasch analysis to determine the measurement properties of a prostate cancer-specific PROM (21) and to evaluate its suitability to collect PROMs for use in clinical practice at the level of individual patients. It has allowed us to scrutinise each aspect of the questionnaire and to identify carefully which aspects work well and which do not.

In our study, the questionnaire was completed at home rather than in clinic and there may be differences between our setting and the setting that was originally used to developed the instrument, especially with respect to the amount of support men received whilst completing the questionnaire.

We also used a different time frame and did not adapt the questions to UK English (as we wanted to evaluate the original questionnaire in its US wording). Yet, it is likely that the anomalies identified in relation to item misfit and inconsistent threshold ordering reflected ambiguous and confusing wording rather than simply linguistic differences between US and UK English.

CONCLUSION

Using Rasch analysis as a diagnostic tool, we have identified several shortcomings of the STAR instrument. In their current form both the urinary function and the sexual function scales have issues that need to be resolved before STAR can be used with confidence in clinical practice. The sexual function scale in particular is unlikely to provide precise estimates for the outcomes experienced by men after radical prostatectomy. For both scales, the underlying construct is not clear and needs further investigation.

Our results demonstrate the need to evaluate the suitability of any PROMs in routine clinical practice, including for example the EPIC-26 that is currently being implemented in prostate cancer care in the UK (10, 11), using modern psychometric methods to identify and address deficiencies that affect their psychometric performance.

Without appropriate psychometric scrutiny and related further development where needed, the use of PROMs in routine clinical practice may significantly misrepresent the true clinical outcomes for patients. PROMs that produce inaccurate and imprecise scores have limited value for clinicians who aim to respond to the needs of their patients. Inaccurate and imprecise scores will also undermine the guiding role that PROMs can have for patients who want to contribute to the management of their own condition. Without progress in development in this area we lose the opportunity to demonstrate the benefit of new technology. This will be detrimental to patients both now and in the future.

Acknowledgements

This work is part of the TrueNTH UK Post Surgery project. TrueNTH is a global initiative by the Movember Foundation, which aims to identify and demonstrate the best and most cost-effective models for improving prostate cancer survivorship care and support. We also thank Dr Jolijn Hendriks for helpful comments on a draft of the manuscript

Author contributorship

EP wrote the first draft of the paper and SS and EP were responsible for the psychometric analysis. CM and JvdM were responsible for the design of the study. All authors contributed to drafting the manuscript and have approved the final version.

Competing Interests Statement

The authors have no conflicts of interest relevant to this article to disclose.

Ethics Statement

Ethical approval for the study was obtained (Study Title: True NTH UK – Post Surgical Follow-up; REC Reference 15/SC/0451).

PPI Statement

Patients and the public were not involved in the design, conduct or dissemination of the project, except as participants in the study.

Funding Statement

This work is funded by Prostate Cancer UK. The funder had no role in any of the following: design and conduct of the study, data collection and management, data analysis and interpretation, or preparation, approval and review of the manuscript.

Financial Disclosure

The authors have no financial relationships relevant to this article to disclose.

References

1. Black N. Patient reported outcome measures could help transform healthcare. BMJ : British Medical Journal. 2013;346.

2. England N. Patient Reported Outcome Measures (PROMs) 2017 [cited 2017 Oct 2017]. Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/proms/.

3. Wales N. Patient Reported Outcome Measures 2017 [cited 2017 Oct 2017]. Available from: https://proms.nhs.wales/.

4. Chard J, Kuczawski M, Black N, van der Meulen J. Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. BMJ. 2011;343.

5. Browne J, Jamieson L, Lewsey J, van der Meulen J, Black N, Cairns J, et al. Patient reported outcome measures (PROMs) in elective surgery. Report to the Department of Health. 2007;12.

6. Baumhauer JF, Bozic KJ. Value-based Healthcare: Patient-reported Outcomes in Clinical Decision Making. Clinical Orthopaedics and Related Research®. 2016;474(6):1375-8.

7. Jason B. Liu M, Andrea L. Pusic, MD, MHS, FACS, Larissa K. Temple, MD, MSc, FACS and Clifford Y. Ko, MD, MS, MSHS, FACS, FASCRS. Patient-reported outcomes in surgery: Listening to patients improves quality of care: Bulletin of the American College of Surgeons; 2017 [Oct 2017]. Available from: http://bulletin.facs.org/2017/03/patient-reported-outcomes-in-surgery-listening-to-patients-improves-quality-of-care/.

8. Vickers AJ, Savage CJ, Shouery M, Eastham JA, Scardino PT, Basch EM. Validation study of a web-based assessment of functional recovery after radical prostatectomy. Health and Quality of Life Outcomes. 2010;8:82.

9. Brundage MD, Barbera L, McCallum F, Howell DM. A pilot evaluation of the expanded prostate cancer index composite for clinical practice (EPIC-CP) tool in Ontario. Qual Life Res. 2018 Oct 31. doi: 10.1007/s11136-018-2034-x. [Epub ahead of print] PubMed PMID: 30382479.

10. Madaan S, Reekhaye A, McFarlane J. Survivorship and prostate cancer: the TrueNTH supported self-management programme. Trends in Urology & Men's Health, January/February 2016:21-24. https://cdn.movember.com/uploads/files/Our%20Work/truenth-supported-self-management-programme-movember-foundation.pdf

11. TrueNTH, a Movember initiative https://prostatecanceruk.org/for-health-professionals/our-projects/truenth

12. US Food and Drug Administration. Guidance for industry on patient-reported outcome measures: Use in medicinal product development to support labeling claims. 2009.

13. Chassany O, Sagnier P, Marquis P, Fullerton S, Aaronson N, Group ERIoQoLA. Patient-reported outcomes: the example of health-related quality of life—a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. Drug Information Journal. 2002;36(1):209-38.

14. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research. 2002;11(3):193-205.

15. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures

used in patient-centered outcomes and comparative effectiveness research. Qual Life Res. 2013;22(8):1889-905.

16. 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):200.

17. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. The Lancet Neurology. 2007;6(12):1094-105.

18. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Press M, editor: MESA Press; 1960.

19. Wright BD, G. M. Rating scale analysis: Rasch measurement. Chicago: MESA; 1982.

20. 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;57(8):1358-62.

21. Protopapa E, van der Meulen J, Moore CM, Smith SC. Patient-reported outcome (PRO) questionnaires for men who have radical surgery for prostate cancer: a conceptual review of existing instruments. BJU international. 2017;120(4):468-81.

22. Wright B. Rack and stack: time 1 vs. time 2. Rasch measurement transactions. 2003;17(1):905-6.

23. Andrich D, Luo G, BE. S. Interpreting RUMM2020. Perth, WA: RUMM Laboratory2004.

24. Andrich D, Sheridan B. RUMM2030. Perth, WA: RUMM Laboratory Pty Ltd; 1997-2017.

25. Iramaneerat, C., Smith Jr., E. & Smith, R. (2008). An introduction to rasch measurement. In Osborne, J. Best practices in quantitative methods (pp. 50-70). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412995627

Table 1: Sample characteristics of the 403 patients who completed at least one valid questionnaire

Sample characteristics		N (%)
Age		
<60		123 (30.5)
60-66		131 (32.5)
>66		149 (37.0)
Ethnicity		
White/White British		321 (79.6)
Other ethnicity		45 (11.2)
Missing		37 (9.2)
Relationship		
Married or living with a partner		309 (76.7)
Other		55 (13.6)
Missing		39 (9.7)
No. of co-morbidities		
0		133 (33.0)
1		164 (40.7)
>2		69 (17.1)
Missing		39 (9.2)

Table 2: Urinary function & sexual function – item fit

Urinary function Item	Location	SE	FitResid	DF	ChiSq	DF	Prob	ICC
Q1 non-complete emptying	-0.492	0.053	-3.077	294.78	15.691	8	0.047026	
Q2 urinate again less than 2hours	0.33	0.035	0.21	598.61	6.623	9	0.676341	
Q3 stopped & started again	-0.329	0.05	-0.591	293.95	8.401	8	0.39534	
Q4 difficult to postpone	-0.155	0.033	2.151	595.31	14.137	9	0.117529	
Q5 weak stream	0.093	0.045	0.499	293.95	7.733	8	0.460021	
Q6 push /strain to begin	-1.103	0.068	-1.731	294.78	6.196	8	0.625333	
Q7 get up in night to urinate	0.238	0.054	3.228	295.6	22.219	8	0.004526	
Q19 leaked urine	0.908	0.047	-1.496	303.83	7.676	9	0.567147	
Q21 pads per day	0.224	0.062	-2.185	304.66	25.356	9	0.002602	
Q23 urinary function - problem	0.287	0.054	-3.157	300.54	27.97	9	0.000965	Questionable
Sexual function Item	**Location**	**SE**	**FitResid**	**DF**	**ChiSq**	**DF**	**Prob**	**ICC**
Q9 confidence to get & keep erection	0.119	0.056	5.814	400.73	135.786	8	0	Questionable
Q10 erection during sexual activity	-0.496	0.049	-2.28	399.9	30.792	8	0.000153	
Q11 erections hard enough for penetration	-0.266	0.05	-3.729	399.08	48.484	8	0	Questionable
Q12 able to penetrate partner	0.195	0.052	-6.208	397.43	49.952	8	0	Questionable
Q13 maintain erection after penetration	0.32	0.053	-5.078	396.61	41.819	8	0.000001	Questionable
Q14 maintain erection to completion	0.129	0.05	-5.152	398.25	35.149	8	0.000025	Questionable

Highlighted items fail criteria

Figure 1a: Urinary Function Person-Item Distribution (targeting)

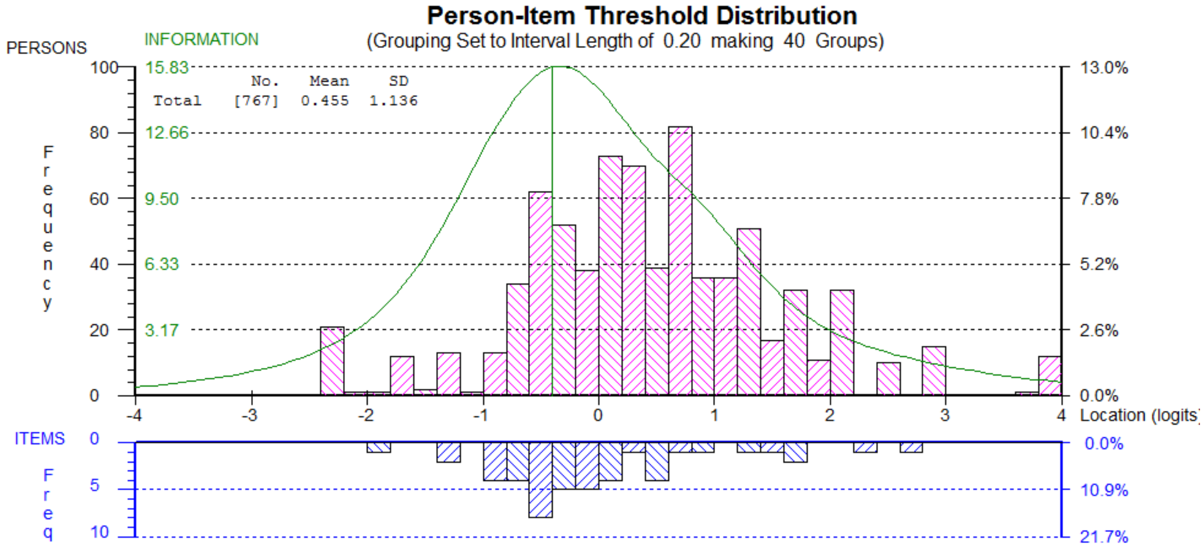
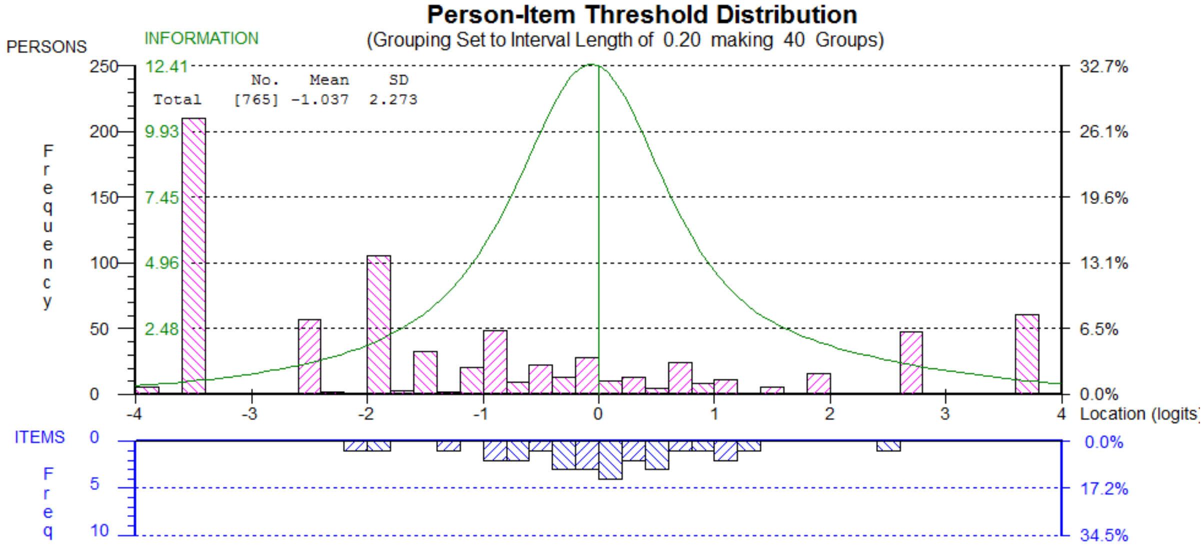
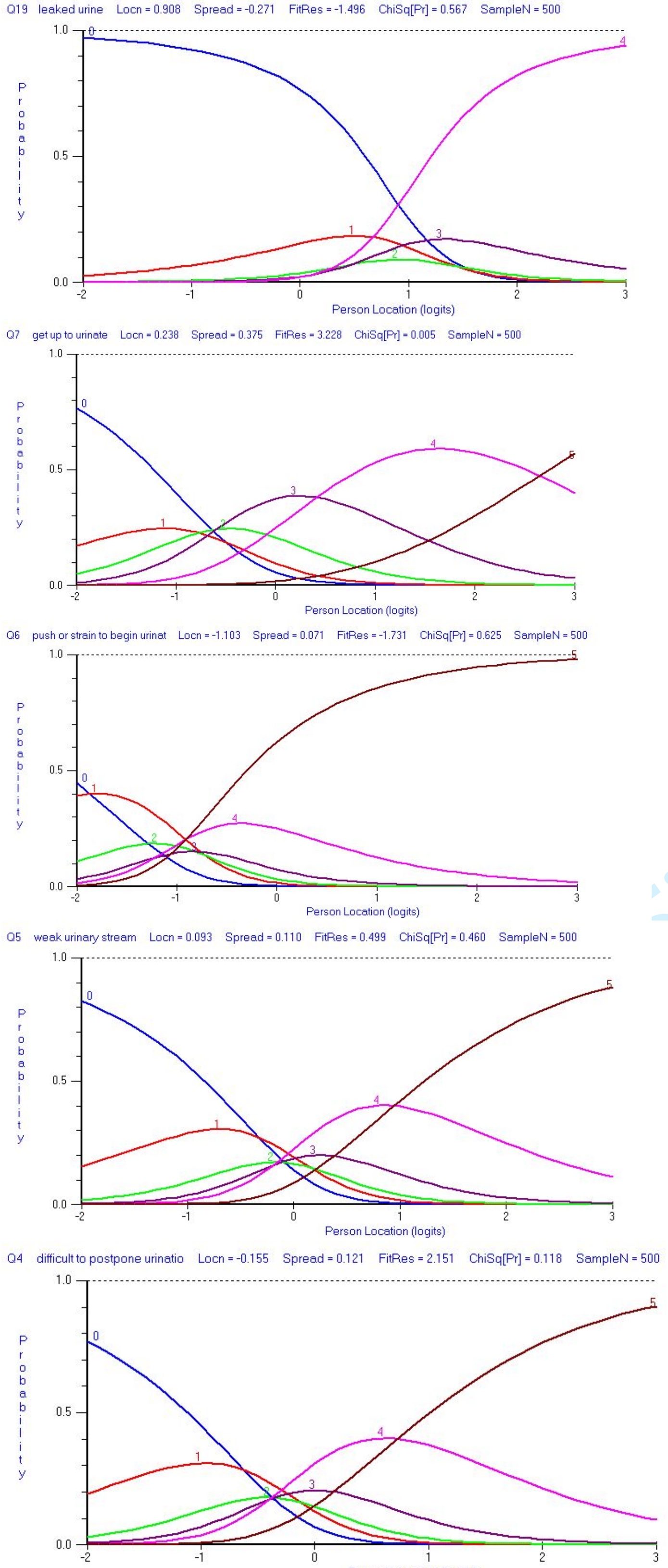


Figure 1b: Sexual Function Person-Item Distribution (targeting)



Figures 2a-2g: Urinary Function Category Probability Curves for disordered items

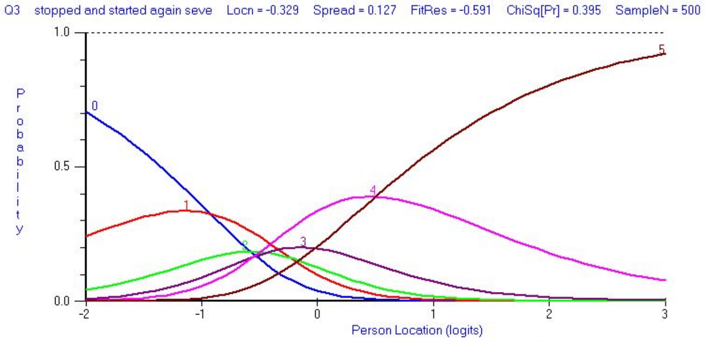

Q19 leaked urine Locn = 0.908 Spread = -0.271 FitRes = -1.496 ChiSq[Pr] = 0.567 SampleN = 500

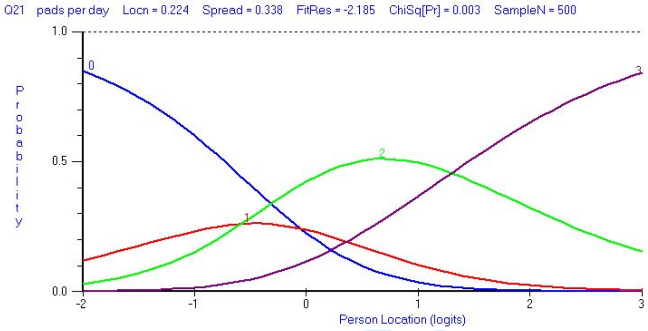
Q7 get up to urinate Locn = 0.238 Spread = 0.375 FitRes = 3.228 ChiSq[Pr] = 0.005 SampleN = 500

Q6 push or strain to begin urinat Locn = -1.103 Spread = 0.071 FitRes = -1.731 ChiSq[Pr] = 0.625 SampleN = 500

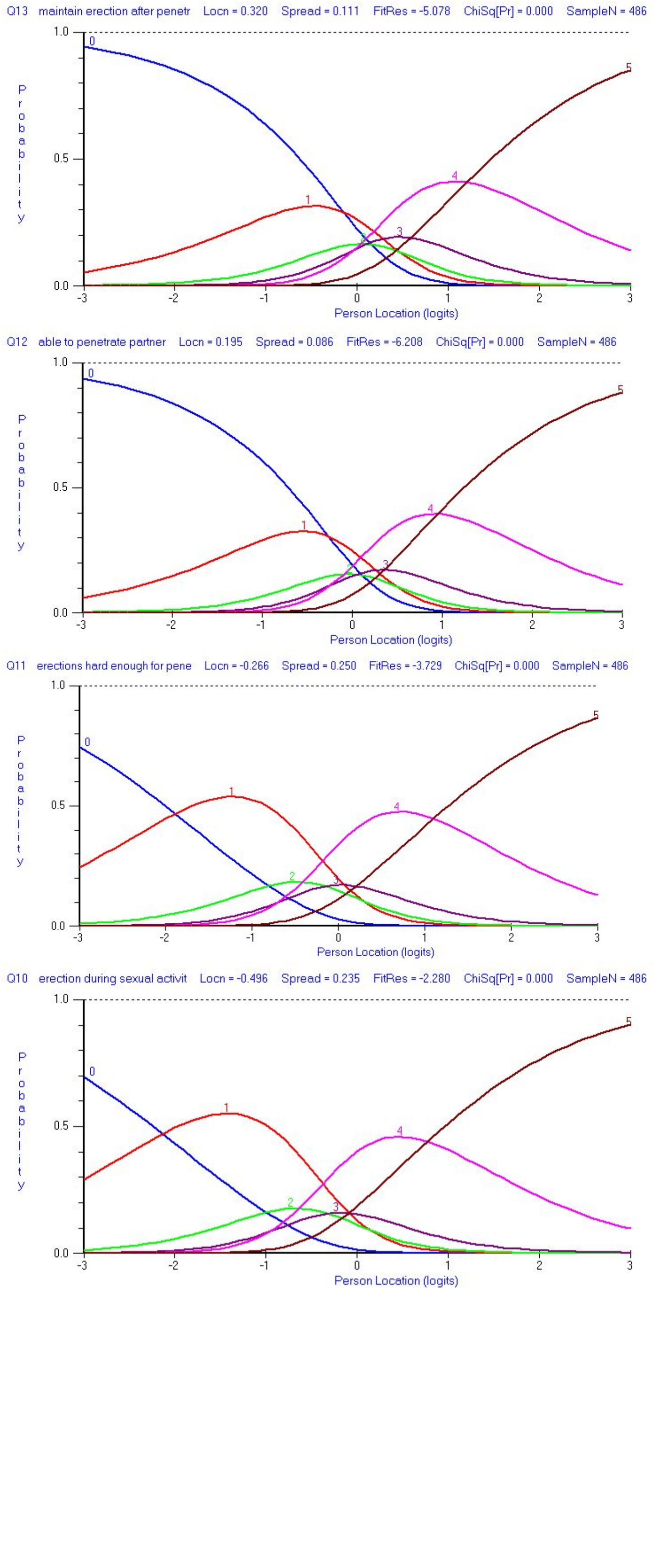
Q5 weak urinary stream Locn = 0.093 Spread = 0.110 FitRes = 0.499 ChiSq[Pr] = 0.460 SampleN = 500

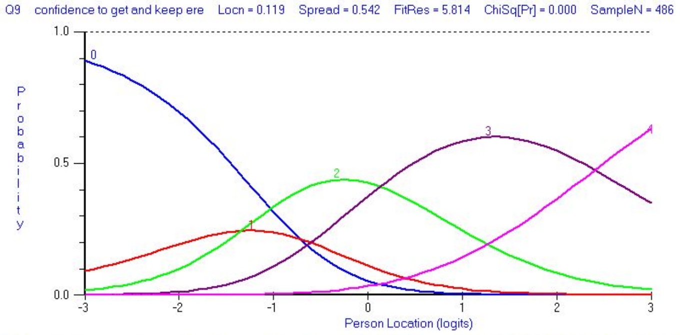
Q4 difficult to postpone urinatio Locn = -0.155 Spread = 0.121 FitRes = 2.151 ChiSq[Pr] = 0.118 SampleN = 500

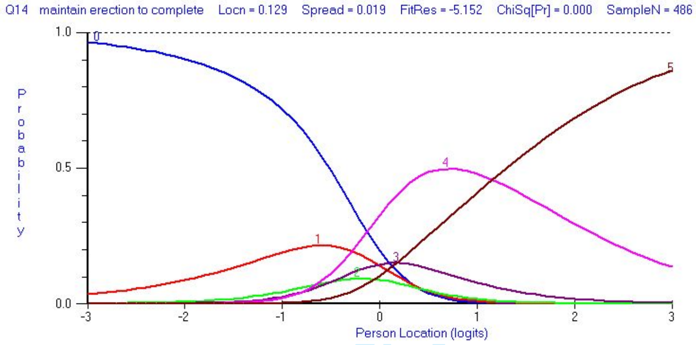

Q3 stopped and started again seve Locn = -0.329 Spread = 0.127 FitRes = -0.591 ChiSq[Pr] = 0.395 SampleN = 500
1.0
0.5
0.0
Probability
0
1
2
3
4
5
-2
-1
0
1
2
3
Person Location (logits)


Q21 pads per day Locn = 0.224 Spread = 0.338 FitRes = -2.185 ChiSq[Pr] = 0.003 SampleN = 500
1.0
0.5
0.0
Probability
0
1
2
3
-2
-1
0
1
2
3
Person Location (logits)

Figures 3a-3f: Sexual Function Category Probability Curves for disordered items




Q9 confidence to get and keep ere Locn = 0.119 Spread = 0.542 FitRes = 5.814 ChiSq[Pr] = 0.000 SampleN = 486
Probability
1.0
0.5
0.0
0
1
2
3
4
-3
-2
-1
0
1
2
3
Person Location (logits)


Q14 maintain erection to complete Locn = 0.129 Spread = 0.019 FitRes = -5.152 ChiSq[Pr] = 0.000 SampleN = 486
Probability
1.0
0.5
0.0
0
1
2
3
4
5
-3
-2
-1
0
1
2
3
Person Location (logits)

BMJ Open

Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-035436.R1
Article Type:	Original research
Date Submitted by the Author:	13-Mar-2020
Complete List of Authors:	Protopapa, Eva; London School of Hygiene and Tropical Medicine, Department of Health Services Research & Policy van der Meulen, Jan; London School of Hygiene and Tropical Medicine, Department of Health Services Research & Policy Moore, Caroline; University College London, Division of Surgery and Interventional Sciences Smith, Sarah; London School of Hygiene & Tropical Medicine, HSRP
<b>Primary Subject Heading</b>:	Urology
Secondary Subject Heading:	Surgery
Keywords:	UROLOGY, Prostate disease < UROLOGY, SURGERY

Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis

Evangelina Protopapa[1] Jan van der Meulen,[1] Caroline M Moore,[2] Sarah Smith[1]

1 London School of Hygiene and Tropical Medicine, London UK

2 University College London, London, UK

Author email addresses:

e.protopapa@ucl.ac.uk

JanvanderMeulen@lshtm.ac.uk

caroline.moore@ucl.ac.uk

Sarah.Smith@lshtm.ac.uk

Address for correspondence;

Dr Sarah Smith, Associate Professor in Psychology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine 15-17 Tavistock Place, London WC1H 9SH. Tel: 0207 9272038

Email: sarah.smith@lshtm.ac.uk

Word count:

Abstract: 299

Text: 3978

Keywords:
PROM, psychometric, Rasch, prostate, surgery

ABSTRACT

OBJECTIVES: Rasch analysis to evaluate the psychometric properties (and identify specific anomalies to be resolved) of urinary and sexual function scales of the STAR instrument for use in clinical practice with individual men.

DESIGN: Prospective cohort study

SETTING: 9 UK surgery centres in secondary care

PARTICIPANTS: 403 men diagnosed with prostate cancer and completed at least one questionnaire immediately before and at 1 or 3 months after a radical prostatectomy.

INTERVENTIONS: Radical prostatectomy.

PRIMARY AND SECONDARY OUTCOMES: STAR instrument before surgery and 1 and 3 months afterwards.

RESULTS: Neither scale fitted the Rasch model (both scales p<0.001). Both urinary (7 items) and sexual function (6 items) had disordered thresholds, suggesting response categories are not working as intendedBoth scales (3 urinary items; 5 sexual function items) showed problems with item fit (large fit residuals, significant chi square,inspection of item characteristic curves (ICC)). Both scales showed items that were unstable over time (DIF by time). Both scales (4 pairs of items in each scale) showed local response dependency (residual correlations >0.2 above the average). Internal consistency was acceptable at the group level for both scales. Targeting was poor for both scales, indicating an inadequate match between location of items and the distribution of the patients, suggesting that the underlying constructs that the scales purport to measure are not clear.

CONCLUSION: Using Rasch analysis as a diagnostic tool, we identified that both the urinary and the sexual function scales have issues that need to be resolved before STAR can be used with confidence in clinical practice. The sexual function scale in particular is unlikely to provide precise estimates for the outcomes experienced by men after radical prostatectomy. These results demonstrate the need to evaluate the suitability of any PROM before implementation in routine clinical practice, preferably using modern psychometric methods**.**

STRENGTHS AND LIMITATIONS OF THIS STUDY

- used modern psychometric methods (based on Rasch measurement Theory) to determine if it is appropriate to use a total function score to describe a patient's sexual or urinary function.
- determined how well the items in each score reflect the experience of men who report the questionnaire
- determined specific anomalies in the scores that suggest that the scales are not being used and understood in the way that was intended
- did not change the items in the questionnaire based on our findings and so did not evaluate any potential improvement such changes would make

INTRODUCTION

The use of patient-reported outcome measures (PROMs) has rapidly increased (1-3). In the UK, PROMs are routinely collected for several areas of elective surgery to evaluate the outcomes in *groups of patients*, receiving a particular treatment or treated in a specific hospital (4, 5). Similar approaches are under consideration for other conditions.

However, there is a lack of evidence about the extent to which clinicians can use PROMs to make their clinical practice more responsive to *individual patients'* needs. Also, it has been suggested that PROMs can play an important role for patients as they can help to inform ways in which patients can self-manage their condition (6, 7).

A web-based tool known as STAR ('Symptom Tracking and Reporting') (8) has been developed at the Memorial Sloan-Kettering Cancer Center (New York, US) to monitor outcomes of radical prostate cancer treatment in individual patients. This

instrument is used to inform both surgeons and men about functional outcomes after surgery, such as urinary, sexual and bowel function improvement or deterioration. Its development is just one example of the implementation of PROMs in prostate cancer practice to inform both clinicians and patients (9-11).

The STAR instrument was not designed to compare men's functional status before and after surgery because different questions are included in the pre- and post-treatment STAR questionnaires. This means that the assessment before surgery is on a different 'ruler' compared to after surgery and therefore there is no clear way of understanding what the change means. However in practice, for example in the English national PROMs programme, pre- and post-treatment PROMs are often compared to monitor the impact of elective surgery (2).

Instruments such as STAR aim to measure specific 'constructs'. It is important these instruments have adequate psychometric properties, otherwise they may produce scores that are 'inaccurate' (prone to systematic error) or 'imprecise' (prone to random error), making it difficult to understand what the observed scores mean and even more difficult to interpret changes over time.

The criteria that must be met to ensure that PROMs are robust are well established (12-15). They ensure that the 'scale' that results from adding up responses to individual questions ('items') relates to a clear underlying construct, as distinct from descriptive responses or simple counts of how many times a symptom occurs.

Like most health-related PROMs, the STAR instrument has been developed using traditional psychometric methods based on classical test theory (CTT). There are important limitations to these methods (16). First, the scales developed using CTT produce 'ordinal scores', where the difference between two adjacent scores at different points on the scale may not be equal. This poses a problem because most statistical analyses assume scores have interval properties where differences between adjacent scores are equal across the entire scale. When scales are based on ordinal scores, changes over time are especially difficult to interpret. Second, the scores can only be interpreted for groups of patients, because measures of statistical uncertainty of these scores (e.g. 'standard errors') are only computed at group level,

which limits their use for individual patients (17). Third, the performance of scales is dependent on the particular sample in which they are used. This makes it difficult to compare studies and, even more importantly, undermines further the interpretation of changes over time.

Modern psychometric methods, such as those based on ‘item-response theory’ (IRT) or ‘Rasch measurement theory’, provide a way of overcoming these challenges. Both are mathematical modelling approaches transforming ordinal scales into interval measures, provided that certain model-related criteria are met. But whereas IRT takes a statistical approach of adding parameters to the model in order to improve its fit to the data, the Rasch paradigm takes a theory-driven approach that investigates why the data do not fit the Rasch model (18-20). The Rasch paradigm, however, keeps central the conceptual underpinning of the instrument and provides a clear set of diagnostic statistics that can help to identify anomalies in its scores.

Instruments developed using these modern psychometric methods have four main advantages over CTT-based instruments. First, they have the potential to generate truly interval scores, thus improving the accuracy and precision with which change over time can be evaluated. Second, measures of statistical uncertainty can be estimated for scores of individual respondents, meaning that the interpretation of scores at patient level is more meaningful. Third, it is possible to produce scales that do not depend on a particular sample’s characteristics. Fourth, they can create a model that contains both pre-and post-surgery items, and therefore all items can be calibrated on the same ruler. The usual pre and post-treatment scores can still be derived but calibrated in such a way that they can be properly compared.

In a systematic review of seven prostate cancer-specific PROMs, including the STAR instrument (21), we identified that modern psychometric methods had not been used to evaluate the psychometric properties of these instruments. In this study, we therefore used Rasch analysis to estimate urinary and sexual function for individual men based on responses to the STAR instrument that were provided by men immediately before and up to three months after radical prostate cancer surgery. In so doing, we identified anomalies that should be addressed to make the

STAR instrument, or any other PROM that aims to monitor changes in outcomes over time after prostate cancer surgery, suitable for use in routine clinical practice.

We performed analyses based on Rasch measurement theory to determine if it is appropriate to use a total function score to describe a patient's sexual or urinary function. As comparisons are often made between pre- and post-surgery scores, we aimed to determine if the seven pre-surgery and five post-surgery items could be placed on the same ruler. If they can then meaningful comparisons can be made across time. To do this, we 'stacked' the data, in other words, we added the baseline and follow-up scores for each patient as separate records (22).

The analyses aimed to answer a number of questions. First, has a measurement ruler been successfully constructed? Second, have the people been successfully measured? Third, is the scale-to-sample targeting adequate? The approach to each of these questions is explained briefly below. A more extensive explanation of Rasch measurement theory can be found in a number of recent overviews (23).

METHODS

Setting and participants

Participants were recruited between November 2015 and March 2017 from nine centres that perform radical prostatectomy by any method (open, laparoscopic-assisted or robotic-assisted) in the UK. Men were eligible if they were diagnosed with prostate cancer, scheduled to have a radical prostatectomy, and had sufficient English language to understand the information about the study and complete the required online questionnaire.

The clinical team at each centre identified and approached eligible patients, informed them about the study, and registered those who were interested in taking part on the secure online portal. Registered patients received their login details by text or email and logged on to the portal to complete the consent form. Once patients had consented, they were directed to the online questionnaire. Patients were invited to

complete the questionnaire before surgery, and at one, three, six and 12 months after surgery.

Instrument

The STAR instrument consists of four domains: sexual function, urinary function, bowel function, and overall quality of life. Our analysis focused on the urinary and sexual function scales obtained immediately before and one and three months after surgery. We excluded the bowel scale from psychometric analyses as with only two items it had insufficient content to be considered a scale. Likewise, the single-item scale for overall quality of life was not considered in our analysis.

Urinary and sexual function items are scored on 3 to 11-point Likert scales. The pre-surgery form of the STAR instrument includes seven urinary function items and the post-surgery form includes five (questions 2 and 4 are common to both). For sexual function, the same six items are included in both pre- and post-surgery forms. Item scores are summed for the urinary and sexual function domains and then transformed to scores ranging from 0 to 100.

We made two wording changes to the STAR instrument. First, our data collection also included the EPIC-26 questionnaire (not reported in the present paper) which overlaps with some STAR items. Where an item existed in both questionnaires, we used the EPIC wording. These minor wording changes are unlikely to substantially change the performance of the item. Second, the standard updated version of STAR has a time frame of six months pre-operatively for both sexual and urinary function, four weeks post operatively for sexual function and one week post operatively for urinary function. To ensure consistency across time for both urinary and sexual function domains, we used a 4-week recall period throughout. We considered this long enough for all problems to be noticed and/or resolved. All items were administered at all time points.

,

Data analysis

Overall fit to the model: For each scale, we evaluated whether the observed responses were significantly different to the responses expected Based on the Rasch model (significant chi square statistic).

Item threshold ordering: For a higher level of functioning on each item, the probability of 'endorsing' a higher response category (on the Likert scale) should increase and the probability of endorsing a lower response category decrease. If each response category in turn (0, 1, 2, 3, 4, 5) has the highest probability of endorsement with increasing levels of functioning, the 'thresholds' between the categories (0-1, 1-2, 2-3, 3-4, 4-5) show a logical order. Thresholds are the location on the scale where the two adjacent response categories have equal probability (50%) of endorsement.

Empirically, however, thresholds can be disordered (e.g. 0-1, 2-3, 1-2), indicating that the response categories do not work as intended. This can be because an item has ambiguous wording or has labels on the response scale that are not sufficiently distinct. We evaluated whether the response categories are working as intended by a visual inspection of the 'category probability curves'.

Item fit validity: The items of the scale must work together ('fit') as a conformable set both clinically and statistically. Clinically, the item ordering along the continuum should make sense and statistically the items need to satisfy specified criteria. Otherwise, it is inappropriate to sum item responses to reach a total score and consider the total score as a measure of the construct. When items do not work together ('misfit') in this way, the validity of a scale needs to be questioned.

We evaluated the fit of each item to the Rasch model by inspecting its 'fit residual' (acceptable range of +/- 2.5) and considering the related Chi-square value. We also assessed visually how closely the observed 'class interval mean scores' follow the expected values in the 'item characteristic curve'. Class intervals are groupings of approximately equal numbers of respondents who have about the same level of functioning.

Differential item functioning (DIF): Stability of the item locations is assessed by evaluating 'differential item functioning (DIF)'. DIF occurs when different groups

within the sample, for example patients of different age, respond differently to an item, despite having the same level of functioning. DIF is identified through an ANOVA main effect for 'person factors', for example age by an interaction between the person factor and the class intervals.

In both the urinary and sexual function scales, we evaluated DIF by age, ethnicity, relationship status and number of co-morbidities. For items that were scored both before and after surgery (two items for the urinary function scale and all six items for the sexual functioning scale), we also evaluated DIF by time point.

Local response dependency: The response to one item should not directly influence the response to another. If 'item response-dependency' happens, measurement estimates can be biased, and reliability, indicated by the 'person separation index', is artificially increased. Local response dependency is evaluated by examining the residual correlations between the items after the Rasch factor they have in common has been partialled out. A correlation coefficient with a value larger than 0.20 above the average of all the item residual correlations indicates potential local response dependency (24).

Reliability: Reliability was examined using the 'person separation index' which is a statistic comparable to the Cronbach's alpha, often used in traditional methods based on CTT. It quantifies how reliably the scale distinguishes between respondents. It is computed from the variation among person locations relative to the standard error of estimate for each individual respondent (16). Higher person separation index values indicate better reliability; a value >0.70 at group level and >0.85 at individual level indicates adequate reliability (20).

Scale to sample targeting: 'Scale-to-sample targeting' describes the match between the range of the construct measured by the items and the range of the construct in the sample of patients. This is evaluated by the 'person-item distribution' which compares the difference between 'person locations' and 'item threshold locations' on the underlying ruler, that captures for example urinary or sexual function. Any gaps in item threshold locations, in particular at the low and high ends of the scale, means

that the functioning of respondents located in that gap area cannot be measured precisely. In other words, their scores will have a relatively large standard error of measurement, because their estimation is severely affected by missing information.

All p-values were adjusted for sample size (n=500) as Chi-square values are sensitive to sample size (25)). Furthermore, Bonferroni corrections for multiple testing were also applied. All analyses were carried using RUMM 2030 (26).

RESULTS

Study sample

Overall, 971 men were eligible, of whom 873 were approached, 714 were interested and 431 men completed the online consent form, giving an overall recruitment rate of 44.4%.

Of the 431 patients who provided consent, 403 patients (93.5%) completed at least one valid questionnaire. A total of 366 valid questionnaires were completed at baseline, 222 questionnaires were completed at one month after surgery and 181 questionnaires at three months after surgery. Table 1 describes the characteristics of the 403 patients included in this analysis. These patients had a mean age of 63 years (SD 6.7; range 41 – 78 years), were predominantly white or white-British (79.7%), and were mostly married or living with a partner (76.7%).

Overall fit to the model

The overall Chi-square statistic indicated that neither the urinary function nor the sexual function scale fit the Rasch model (urinary function, chi square=207.04p<0.001; sexual function, chi square=341.98; p<0.001).

Item threshold ordering

Both urinary and sexual function scales had items with disordered thresholds, indicating that the response options were not working as intended. The urinary function scale had disordered thresholds for 7 of the 10 items. For these 7 disordered items, the category probability plots in Figures 1a-1g illustrate that this is mainly a problem with the middle response options, suggesting that the wording was not clear or that the difference between categories was not well understood. For example, for Q3 of the urinary function scale ('Over the last 4 weeks, how often have you found you stopped and started again several times when you urinated?') there is no point at which threshold 2 ('About half the time') and threshold 3 ('Less than half the time') are the most likely to occur. If the response options were working as intended, the probability of each threshold should come in order.

All six of the sexual function items are disordered. This means that none of the response scales are working as they were intended. Figures 2a-2f indicate that it is mainly thresholds 2 and 3 that are disordered, suggesting that the middle categories of the response scales are not well understood and may need to be re-worded.

Item fit validity

Both the urinary and sexual function scales contained items that did not fit the model, when considering together their fit residual, Chi-square value, and the item characteristic curve (fit residuals and chi Square values for all items are reported in Table 2). One urinary function item (Q23) failed all three criteria indicating misfit to the model. Two further items failed one or two criteria (Q3 and Q7) indicating a broader problem with item fit.

Five sexual function items failed all three criteria (Table 2) and the remaining item failed one of the three criteria suggesting further problems with item fit.

Differential item functioning (DIF)

Overall, items in both scales were stable (invariant) across different groups for age, ethnicity, relationship status and number of co-morbidities. However, both scales contained items that were unstable across time, with the sexual function scale containing a greater number of unstable items.

One urinary item (Q23) showed DIF across time points (p<0.001). Patients' response to this item were systematically higher at 3 months post-op compared to 1 month post-surgery, despite having equal underlying levels of urinary function.

Five sexual function item (Q9, Q10, Q11, Q12, Q13) showed DIF by time (p<0.001)

Local response dependency

Both scales contained pairs of items that were dependent on each other, but the sexual function scale showed greater local dependency. Four pairs of urinary function items showed local dependency: Q3 (stopped and started again) and Q4 (difficulty postponing urination) (residual correlation = 0.10); Q5 (weak urinary stream) and Q6 (push or strain to begin urination) (residual correlation = 0.04); Q19 (leaking urine) and Q21 (number of pads per day) (residual correlation = 0.32); Q21 (number of pads per day) and Q23 (urinary problem overall) (residual correlation = 0.13).

Four pairs of sexual function items showed local dependency with relatively high residual correlations: Q10 (erection during sexual activity) and Q11 (erections hard enough for penetration) (residual correlation = 0.30), Q12 (able to penetrate) and Q13 (maintain erection after penetration) (residual correlation = 0.59), Q12 (able to penetrate) and Q14 (maintain erection to completion) (residual correlation = 0.55), Q13 (maintain erection after penetration) and Q14 (maintain erection to completion) (residual correlation = 0.51).

Reliability

Internal consistency was acceptable at group level for both scales (urinary function scale: person separation index = 0.75; sexual function scale: person separation index = 0.82).

Scale-to-sample targeting

The person-item distribution of the urinary function scale was relatively poor, though better than the targeting for the sexual function scale (Figure 3a). Although the middle of the person distribution is reasonably well matched by items, both extremes of the distribution have few items. This means that for men located at the lower end of the scale (including many men at one month after surgery) and at the higher end of the scale (including many men before surgery) the level of functioning cannot be precisely measured.

The targeting for the sexual functioning scale was also poor (Figure 3b). In particular, most items are located in the centre of the scale whereas the distribution of people is quite wide. This means that the sexual function for men located at the higher end of the scale (often men before surgery) and the lower end of the scale (most of the men after surgery) is very imprecisely measured.

DISCUSSION

Our analyses have identified that neither the urinary function items nor the sexual function items from the STAR instrument can be placed on a common metric that is robust for comparisons before and after surgery. Furthermore, a number of anomalies have been identified that suggest the scales are not working as intended. There is an inadequate match between location of items and the distribution of the patients, suggesting that the underlying constructs that the scales purport to measure are not clear. Consequently, the items do not measure the men's function very accurately. The response categories for many items are not consistently used, some items do not work with the others as a conformable set and some items are not stable over time.

These results indicate that in its current form the items in the STAR instrument do not provide an adequate ruler to monitor urinary or sexual function in clinical practice. These problems are likely to make the estimation of an individual patient's outcome after surgery less accurate and precise and using the questionnaire in its current form therefore carries a risk of misrepresenting actual urinary and sexual outcomes.

Our results demonstrate that the risk of inaccurate estimation of outcomes using STAR is likely to be most pronounced for men with either very good or very poor outcomes. The poor scale-to-sample targeting, particularly for the sexual functioning scale, also means that this problem is exacerbated for men with better function before surgery and worse function after surgery, creating clear problems for the interpretation of change scores that are supposed to capture the impact of surgery. Further, both scales have items that showed DIF by time providing further evidence that it is not meaningful to compare scores before and after surgery or compare scores taken at different times after surgery.

In the short term, some of the identified deficiencies can be addressed using post-hoc statistical techniques to re-score the disordered thresholds (16, 20) or to resolve for the uniform DIF (23) and local response dependency (20). However, a more robust solution would be to conduct qualitative research with men who have experienced radical prostatectomy to understand why the questions are not well understood and why the response options are not used in the way that was intended. Qualitative research should also explore which areas of content are missing and how items could be formulated to address these gaps. A revised version based on these findings would then need to be psychometrically evaluated again to determine how well the amendments to content and scoring have addressed the identified problems.

This study is the first to use robust modern psychometric methods such as Rasch analysis to determine the measurement properties of a prostate cancer-specific PROM (21) and to evaluate its suitability to collect PROMs for use in clinical practice at the level of individual patients. It has allowed us to scrutinise each aspect of the questionnaire and to identify carefully which aspects work well and which do not.

In our study, the questionnaire was completed at home rather than in clinic and there may be differences between our setting and the setting that was originally used to developed the instrument, especially with respect to the amount of support men received whilst completing the questionnaire.

We also used a different time frame and did not adapt the questions to UK English (as we wanted to evaluate the original questionnaire in its US wording). Yet, it is likely that the anomalies identified in relation to item misfit and inconsistent threshold ordering reflected ambiguous and confusing wording rather than simply linguistic differences between US and UK English.

CONCLUSION

Using Rasch analysis as a diagnostic tool, we have identified several shortcomings of the STAR instrument. In their current form both the urinary function and the sexual function scales have issues that need to be resolved before STAR can be used with confidence in clinical practice. The sexual function scale in particular is unlikely to provide precise estimates for the outcomes experienced by men after radical prostatectomy. For both scales, the underlying construct is not clear and needs further investigation.

Our results demonstrate the need to evaluate the suitability of any PROMs in routine clinical practice, including for example the EPIC-26 that is currently being implemented in prostate cancer care in the UK (10, 11), using modern psychometric methods to identify and address deficiencies that affect their psychometric performance.

Without appropriate psychometric scrutiny and related further development where needed, the use of PROMs in routine clinical practice may significantly misrepresent the true clinical outcomes for patients. PROMs that produce inaccurate and imprecise scores have limited value for clinicians who aim to respond to the needs of their patients. Inaccurate and imprecise scores will also undermine the guiding role that PROMs can have for patients who want to contribute to the management of their own condition. Without progress in development in this area we lose the opportunity to demonstrate the benefit of new technology. This will be detrimental to patients both now and in the future.

Acknowledgements

This work is part of the TrueNTH UK Post Surgery project. TrueNTH is a global initiative by the Movember Foundation, which aims to identify and demonstrate the best and most cost-effective models for improving prostate cancer survivorship care and support. We also thank Dr Jolijn Hendriks for helpful comments on a draft of the manuscript

Author contributorship

EP wrote the first draft of the paper and SS and EP were responsible for the psychometric analysis. CM and JvdM were responsible for the design of the study.

All authors contributed to drafting the manuscript and have approved the final version.

Competing Interests Statement

The authors have no conflicts of interest relevant to this article to disclose.

Ethics Statement

Ethical approval for the study was obtained (Study Title: True NTH UK – Post Surgical Follow-up; REC Reference 15/SC/0451).

PPI Statement

Patients and the public were not involved in the design, conduct or dissemination of the project, except as participants in the study.

Funding Statement

This work is funded by Prostate Cancer UK. The funder had no role in any of the following: design and conduct of the study, data collection and management, data analysis and interpretation, or preparation, approval and review of the manuscript.

Financial Disclosure

The authors have no financial relationships relevant to this article to disclose.

Data Availability Statement

No additional data available

References

1. Black N. Patient reported outcome measures could help transform healthcare. BMJ : British Medical Journal. 2013;346.

2. England N. Patient Reported Outcome Measures (PROMs) 2017 [cited 2017 Oct 2017]. Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/proms/.

3. Wales N. Patient Reported Outcome Measures 2017 [cited 2017 Oct 2017]. Available from: https://proms.nhs.wales/.

4. Chard J, Kuczawski M, Black N, van der Meulen J. Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. BMJ. 2011;343.

5. Browne J, Jamieson L, Lewsey J, van der Meulen J, Black N, Cairns J, et al. Patient reported outcome measures (PROMs) in elective surgery. Report to the Department of Health. 2007;12.

6. Baumhauer JF, Bozic KJ. Value-based Healthcare: Patient-reported Outcomes in Clinical Decision Making. Clinical Orthopaedics and Related Research®. 2016;474(6):1375-8.

7. Jason B. Liu M, Andrea L. Pusic, MD, MHS, FACS, Larissa K. Temple, MD, MSc, FACS and Clifford Y. Ko, MD, MS, MSHS, FACS, FASCRS. Patient-reported outcomes in surgery: Listening to patients improves quality of care: Bulletin of the American College of Surgeons; 2017 [Oct 2017]. Available from: http://bulletin.facs.org/2017/03/patient-reported-outcomes-in-surgery-listening-to-patients-improves-quality-of-care/.

8. Vickers AJ, Savage CJ, Shouery M, Eastham JA, Scardino PT, Basch EM. Validation study of a web-based assessment of functional recovery after radical prostatectomy. Health and Quality of Life Outcomes. 2010;8:82.

9. Brundage MD, Barbera L, McCallum F, Howell DM. A pilot evaluation of the expanded prostate cancer index composite for clinical practice (EPIC-CP) tool in Ontario. Qual Life Res. 2018 Oct 31. doi: 10.1007/s11136-018-2034-x. [Epub ahead of print] PubMed PMID: 30382479.

10. Madaan S, Reekhaye A, McFarlane J. Survivorship and prostate cancer: the TrueNTH supported self-management programme. Trends in Urology & Men's Health, January/February 2016:21-24. https://cdn.movember.com/uploads/files/Our%20Work/truenth-supported-self-management-programme-movember-foundation.pdf

11. TrueNTH, a Movember initiative https://prostatecanceruk.org/for-health-professionals/our-projects/truenth

12. US Food and Drug Administration. Guidance for industry on patient-reported outcome measures: Use in medicinal product development to support labeling claims. 2009.

13. Chassany O, Sagnier P, Marquis P, Fullerton S, Aaronson N, Group ERIoQoLA. Patient-reported outcomes: the example of health-related quality of life—a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. Drug Information Journal. 2002;36(1):209-38.

14. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research. 2002;11(3):193-205.

15. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures

used in patient-centered outcomes and comparative effectiveness research. Qual Life Res. 2013;22(8):1889-905.

16. 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):200.

17. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. The Lancet Neurology. 2007;6(12):1094-105.

18. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Press M, editor: MESA Press; 1960.

19. Wright BD, G. M. Rating scale analysis: Rasch measurement. Chicago: MESA; 1982.

20. 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;57(8):1358-62.

21. Protopapa E, van der Meulen J, Moore CM, Smith SC. Patient-reported outcome (PRO) questionnaires for men who have radical surgery for prostate cancer: a conceptual review of existing instruments. BJU international. 2017;120(4):468-81.

22. Wright B. Rack and stack: time 1 vs. time 2. Rasch measurement transactions. 2003;17(1):905-6.

23. Andrich D, Luo G, BE. S. Interpreting RUMM2020. Perth, WA: RUMM Laboratory2004.

24. Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q 3 : Identification of Local Dependence in the Rasch Model Using Residual Correlations. Applied Psychological Measurement, 41(3), 178–194. https://doi.org/10.1177/0146621616677520

25. Andrich D, Sheridan B. RUMM2030. Perth, WA: RUMM Laboratory Pty Ltd; 1997-2017.

26. Iramaneerat, C., Smith Jr., E. & Smith, R. (2008). An introduction to rasch measurement. In Osborne, J. Best practices in quantitative methods (pp. 50-70). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412995627

Table 1: Sample characteristics of the 403 patients who completed at least one valid questionnaire

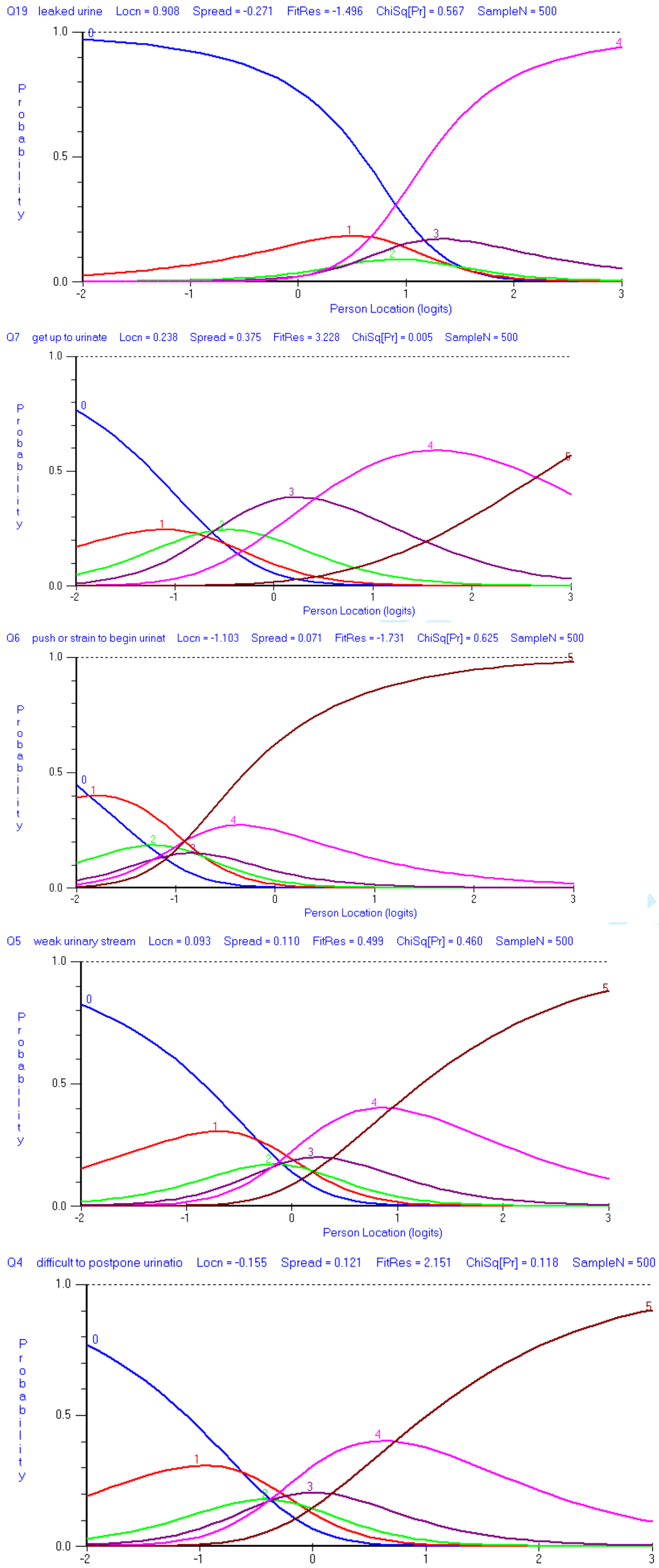
Sample characteristics			**N (%)**
Age			
<60			123 (30.5)
60-66			131 (32.5)
>66			149 (37.0)
Ethnicity			
White/White British			321 (79.6)
Other ethnicity			45 (11.2)
Missing			37 (9.2)
Relationship			
Married or living with a partner			309 (76.7)
Other			55 (13.6)
Missing			39 (9.7)
No. of co-morbidities			
0			133 (33.0)
1			164 (40.7)
>2			69 (17.1)
Missing			39 (9.2)

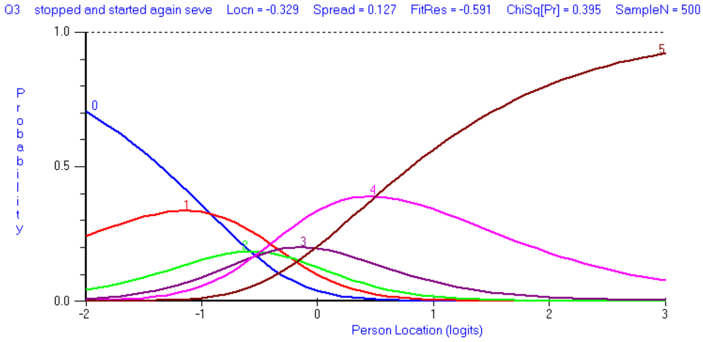
Table 2: Urinary function & sexual function – item fit

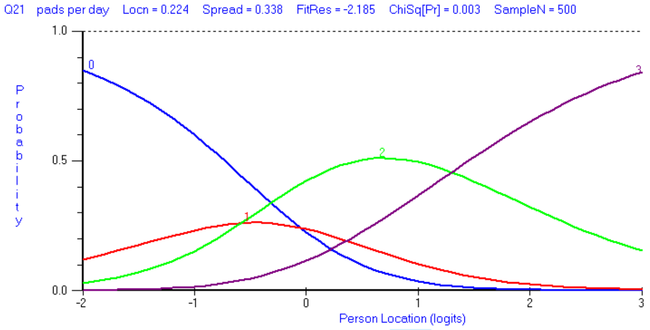
Urinary function Item	Location	SE	FitResid	DF	ChiSq	DF	Prob	ICC
Q1 non-complete emptying	-0.492	0.053	-3.077	294.78	15.691	8	0.047026	
Q2 urinate again less than 2hours	0.33	0.035	0.21	598.61	6.623	9	0.676341	
Q3 stopped & started again	-0.329	0.05	-0.591	293.95	8.401	8	0.39534	
Q4 difficult to postpone	-0.155	0.033	2.151	595.31	14.137	9	0.117529	
Q5 weak stream	0.093	0.045	0.499	293.95	7.733	8	0.460021	
Q6 push /strain to begin	-1.103	0.068	-1.731	294.78	6.196	8	0.625333	
Q7 get up in night to urinate	0.238	0.054	3.228	295.6	22.219	8	0.004526	
Q19 leaked urine	0.908	0.047	-1.496	303.83	7.676	9	0.567147	
Q21 pads per day	0.224	0.062	-2.185	304.66	25.356	9	0.002602	
Q23 urinary function - problem	0.287	0.054	-3.157	300.54	27.97	9	0.000965	Questionable
Sexual function Item	**Location**	**SE**	**FitResid**	**DF**	**ChiSq**	**DF**	**Prob**	**ICC**
Q9 confidence to get & keep erection	0.119	0.056	5.814	400.73	135.786	8	0	Questionable
Q10 erection during sexual activity	-0.496	0.049	-2.28	399.9	30.792	8	0.000153	
Q11 erections hard enough for penetration	-0.266	0.05	-3.729	399.08	48.484	8	0	Questionable
Q12 able to penetrate partner	0.195	0.052	-6.208	397.43	49.952	8	0	Questionable
Q13 maintain erection after penetration	0.32	0.053	-5.078	396.61	41.819	8	0.000001	Questionable
Q14 maintain erection to completion	0.129	0.05	-5.152	398.25	35.149	8	0.000025	Questionable

Highlighted items fail criteria

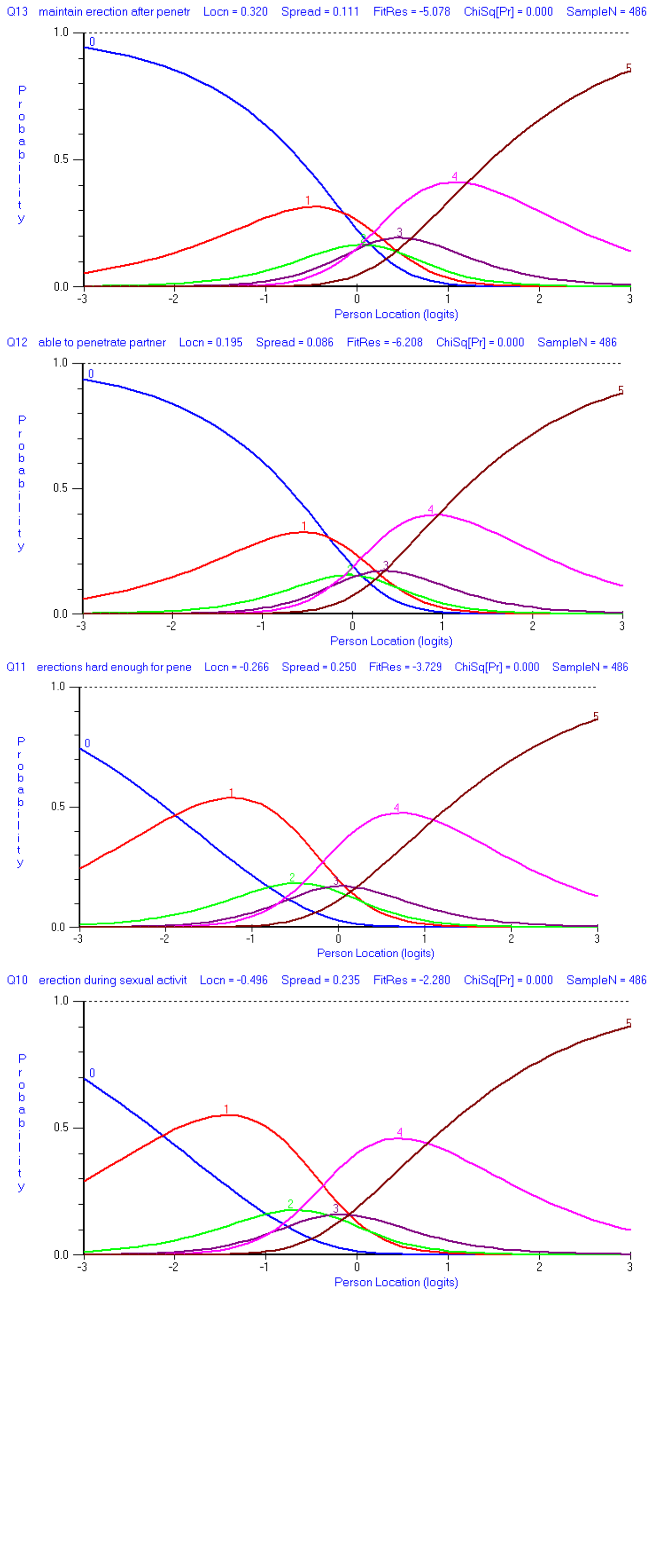
Figures 1a-1g: Urinary Function Category Probability Curves for disordered items




Q3 stopped and started again seve Locn = -0.329 Spread = 0.127 FitRes = -0.591 ChiSq[Pr] = 0.395 SampleN = 500
Probability
1.0
0.5
0.0
-2
-1
0
1
2
3
Person Location (logits)
0
1
2
3
4
5


Q21 pads per day Locn = 0.224 Spread = 0.338 FitRes = -2.185 ChiSq[Pr] = 0.003 SampleN = 500
Probability
1.0
0.5
0.0
-2
-1
0
1
2
3
Person Location (logits)
0
1
2
3

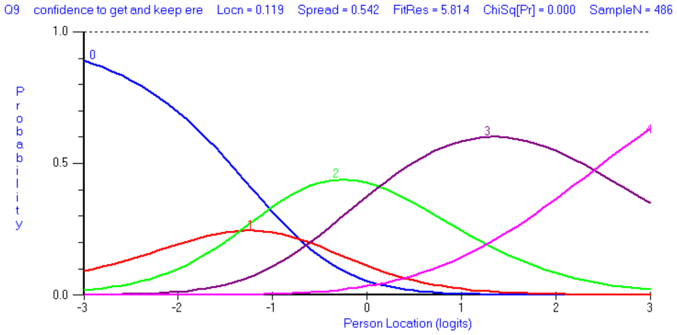
Figures 2a-2f: Sexual Function Category Probability Curves for disordered items


Q13 maintain erection after penetr Locn = 0.320 Spread = 0.111 FitRes = -5.078 ChiSq[Pr] = 0.000 SampleN = 486

Q12 able to penetrate partner Locn = 0.195 Spread = 0.086 FitRes = -6.208 ChiSq[Pr] = 0.000 SampleN = 486

Q11 erections hard enough for pene Locn = -0.266 Spread = 0.250 FitRes = -3.729 ChiSq[Pr] = 0.000 SampleN = 486

Q10 erection during sexual activit Locn = -0.496 Spread = 0.235 FitRes = -2.280 ChiSq[Pr] = 0.000 SampleN = 486


Q9 confidence to get and keep ere Locn = 0.119 Spread = 0.542 FitRes = 5.814 ChiSq[Pr] = 0.000 SampleN = 486
Probability
1.0
0.5
0.0
-3
-2
-1
0
1
2
3
Person Location (logits)
0
1
2
3
4

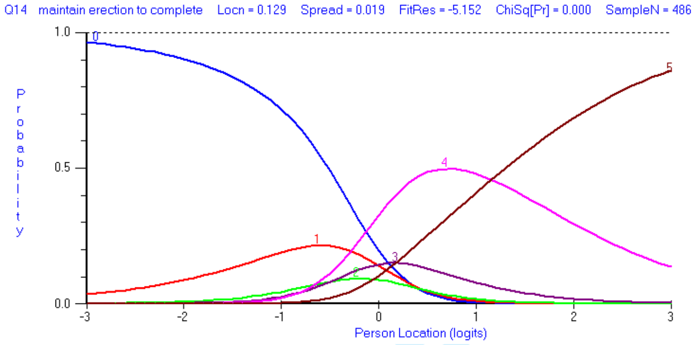

Q14 maintain erection to complete Locn = 0.129 Spread = 0.019 FitRes = -5.152 ChiSq[Pr] = 0.000 SampleN = 486
Probability
1.0
0.5
0.0
-3
-2
-1
0
1
2
3
Person Location (logits)
0
1
3
4
5

Figure 3a: Urinary Function Person-Item Distribution (targeting)

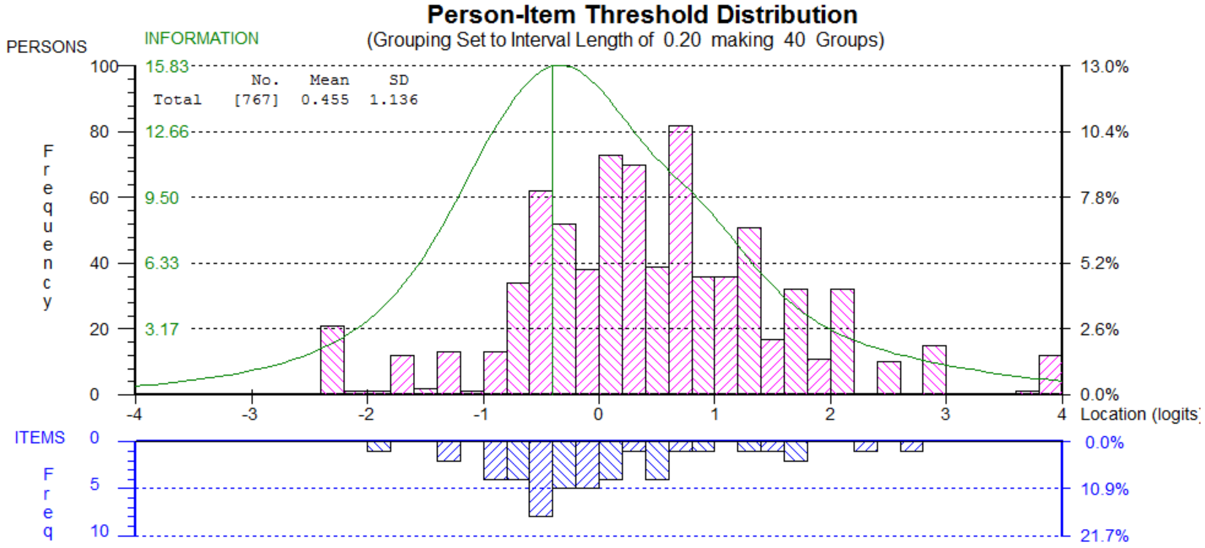
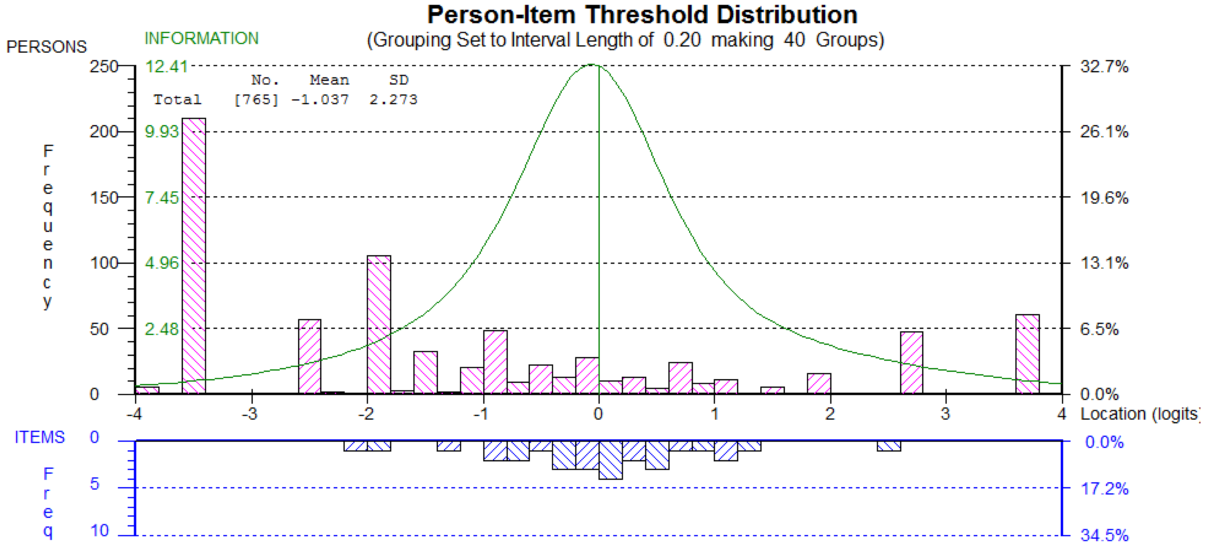


Figure 3b: Sexual Function Person-Item Distribution (targeting)



BMJ Open

Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-035436.R2
Article Type:	Original research
Date Submitted by the Author:	07-May-2020
Complete List of Authors:	Protopapa, Eva; London School of Hygiene and Tropical Medicine, Department of Health Services Research & Policy van der Meulen, Jan; London School of Hygiene and Tropical Medicine, Department of Health Services Research & Policy Moore, Caroline; University College London, Division of Surgery and Interventional Sciences Smith, Sarah; London School of Hygiene & Tropical Medicine, HSRP
<b>Primary Subject Heading</b>:	Urology
Secondary Subject Heading:	Surgery
Keywords:	UROLOGY, Prostate disease < UROLOGY, SURGERY

Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis

Evangelina Protopapa[1] Jan van der Meulen,[1] Caroline M Moore,[2] Sarah Smith[1]

1 London School of Hygiene and Tropical Medicine, London UK
2 University College London, London, UK

Author email addresses:
e.protopapa@ucl.ac.uk
JanvanderMeulen@lshtm.ac.uk
caroline.moore@ucl.ac.uk
Sarah.Smith@lshtm.ac.uk

Address for correspondence;
Dr Sarah Smith, Associate Professor in Psychology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine 15-17 Tavistock Place, London WC1H 9SH. Tel: 0207 9272038
Email: sarah.smith@lshtm.ac.uk

Word count:
Abstract: 300
Text: 4165

Keywords:
PROM, psychometric, Rasch, prostate, surgery

ABSTRACT

OBJECTIVES: To evaluate the psychometric properties (and identify specific anomalies to be resolved) of urinary and sexual function scales of the STAR instrument for use in clinical practice with individual men using Rasch analysis.

DESIGN: Prospective cohort study

SETTING: 9 UK surgery centres in secondary care

PARTICIPANTS: 403 men diagnosed with prostate cancer and completed at least one questionnaire immediately before and at 1 or 3 months after a radical prostatectomy.

INTERVENTIONS: Radical prostatectomy.

PRIMARY AND SECONDARY OUTCOMES: STAR instrument before surgery and 1 and 3 months afterwards.

RESULTS: Neither scale fitted the Rasch model (both scales p<0.001). Both urinary (7 items) and sexual function (6 items) had disordered thresholds, suggesting response categories are not working as intended. Both scales (3 urinary items; 5 sexual function items) showed problems with item fit (large fit residuals, significant chi square, inspection of item characteristic curves (ICC)). Both scales showed items that were unstable over time (DIF by time). Both scales (4 pairs of items in each scale) showed local response dependency (residual correlations >0.2 above the average). Internal consistency was acceptable at the group level for both scales. Targeting was poor for both scales, indicating an inadequate match between location of items and the distribution of the patients, suggesting that the underlying constructs that the scales purport to measure are not clear.

CONCLUSION: Using Rasch analysis as a diagnostic tool, we identified that both the urinary and the sexual function scales have issues that need to be resolved before STAR can be used with confidence in clinical practice. The sexual function scale in particular is unlikely to provide precise estimates for the outcomes experienced by men after radical prostatectomy. These results demonstrate the need to evaluate the suitability of any PROM before implementation in routine clinical practice, preferably using modern psychometric methods**.**

STRENGTHS AND LIMITATIONS OF THIS STUDY

- used modern psychometric methods (based on Rasch measurement Theory) to determine if it is appropriate to use a total function score to describe a patient's sexual or urinary function.
- determined how well the items in each scale reflect the experience of men who report the questionnaire
- determined specific anomalies in the scores that suggest that the scales are not being used and understood in the way that was intended
- did not change the items in the questionnaire based on our findings and so did not evaluate any potential improvement such changes would make

INTRODUCTION

The use of patient-reported outcome measures (PROMs) has rapidly increased (1-3). In the UK, PROMs are routinely collected for several areas of elective surgery to evaluate the outcomes in *groups of patients*, receiving a particular treatment or treated in a specific hospital (4, 5). Similar approaches are under consideration for other conditions.

However, there is a lack of evidence about the extent to which clinicians can use PROMs to make their clinical practice more responsive to *individual patients'* needs. Also, it has been suggested that PROMs can play an important role for patients as they can help to inform ways in which patients can self-manage their condition (6, 7).

A web-based tool known as STAR ('Symptom Tracking and Reporting') (8) has been developed at the Memorial Sloan-Kettering Cancer Center (New York, US) to monitor outcomes of radical prostate cancer treatment in individual patients. This

instrument is used to inform both surgeons and men about functional outcomes after surgery, such as urinary, sexual and bowel function improvement or deterioration. Its development is just one example of the implementation of PROMs in prostate cancer practice to inform both clinicians and patients (9-11).

The STAR instrument was not designed to compare men's functional status before and after surgery because different questions are included in the pre- and post-treatment STAR questionnaires. This means that the assessment before surgery is on a different 'ruler' compared to after surgery and therefore there is no clear way of understanding what the change means. However in practice, for example in the English national PROMs programme, pre- and post-treatment PROMs are often compared to monitor the impact of elective surgery (2).

Instruments such as STAR aim to measure specific 'constructs'. It is important these instruments have adequate psychometric properties, otherwise they may produce scores that are 'inaccurate' (prone to systematic error) or 'imprecise' (prone to random error), making it difficult to understand what the observed scores mean and even more difficult to interpret changes over time.

The criteria that must be met to ensure that PROMs are robust are well established (12-15). They ensure that the 'scale' that results from adding up responses to individual questions ('items') relates to a clear underlying construct, as distinct from descriptive responses or simple counts of how many times a symptom occurs.

Like most health-related PROMs, the STAR instrument has been developed using traditional psychometric methods based on classical test theory (CTT). There are important limitations to these methods (16). First, the scales developed using CTT produce 'ordinal scores', where the difference between two adjacent scores at different points on the scale may not be equal. This poses a problem because most statistical analyses assume scores have interval properties where differences between adjacent scores are equal across the entire scale. When scales are based on ordinal scores, changes over time are especially difficult to interpret. Second, the scores can only be interpreted for groups of patients, because measures of statistical uncertainty of these scores (e.g. 'standard errors') are only computed at group level,

which limits their use for individual patients (17). Third, the performance of scales is dependent on the particular sample in which they are used. This makes it difficult to compare studies and, even more importantly, undermines further the interpretation of changes over time.

Modern psychometric methods, such as those based on ‘item response theory’ (IRT) or ‘Rasch measurement theory’, provide a way of overcoming these challenges. Both are mathematical modelling approaches transforming ordinal scales into interval measures, provided that certain model-related criteria are met. But whereas IRT takes a statistical approach of adding parameters to the model in order to improve its fit to the data, the Rasch paradigm takes a theory-driven approach that investigates why the data do not fit the Rasch model (18-20). The Rasch paradigm, however, keeps central the conceptual underpinning of the instrument and provides a clear set of diagnostic statistics that can help to identify anomalies in its scores.

Instruments developed using these modern psychometric methods have four main advantages over CTT-based instruments. First, they have the potential to generate truly interval scores, thus improving the accuracy and precision with which change over time can be evaluated. Second, measures of statistical uncertainty can be estimated for scores of individual respondents, meaning that the interpretation of scores at patient level is more meaningful. Third, it is possible to produce scales that do not depend on a particular sample’s characteristics. Fourth, they can create a model that contains both pre-and post-surgery items, and therefore all items can be calibrated on the same ruler. The usual pre and post-treatment scores can still be derived but calibrated in such a way that they can be properly compared.

In a systematic review of seven prostate cancer-specific PROMs, including the STAR instrument (21), we identified that modern psychometric methods had not been used to evaluate the psychometric properties of these instruments. In this study, we therefore used Rasch analysis to estimate urinary and sexual function for individual men based on responses to the STAR instrument that were provided by men immediately before and up to three months after radical prostate cancer surgery. In so doing, we aimed to identify anomalies that should be addressed to make the STAR instrument, or any other PROM that aims to monitor changes in outcomes over time after prostate cancer surgery, suitable for use in routine clinical practice.

We performed analyses based on Rasch measurement theory to determine if it is appropriate to use a total function score to describe a patient's sexual or urinary function. As comparisons are often made between pre- and post-surgery scores, we aimed to determine if the seven pre-surgery and five post-surgery items could be placed on the same measurement ruler. If they can, then meaningful comparisons can be made across time. To do this, we 'stacked' the data, in other words, we added the baseline and follow-up scores for each patient as separate records (22).

The analyses aimed to answer a number of questions. First, has a measurement ruler been successfully constructed? Second, have the people been successfully measured? Third, is the scale-to-sample targeting adequate? The approach to each of these questions is explained briefly below. A more extensive explanation of Rasch measurement theory can be found in a recent overviews (23).

METHODS

Setting and participants

Participants were recruited between November 2015 and March 2017 from nine centres that perform radical prostatectomy by any method (open, laparoscopic-assisted or robotic-assisted) in the UK. Men were eligible if they were diagnosed with prostate cancer, scheduled to have a radical prostatectomy, and had sufficient English language to understand the information about the study and complete the required online questionnaire.

The clinical team at each centre identified and approached eligible patients, informed them about the study, and registered those who were interested in taking part on the secure online portal. Registered patients received their login details by text or email and logged on to the portal to complete the consent form. Once patients had consented, they were directed to the online questionnaire. Patients were invited to

complete the questionnaire before surgery, and at one, three, six and 12 months after surgery.

Instrument

The STAR instrument consists of four domains: sexual function, urinary function, bowel function, and overall quality of life. Our analysis focused on the urinary and sexual function scales obtained immediately before and one and three months after surgery. We excluded the bowel scale from psychometric analyses as with only two items it had insufficient content to be considered a scale. Likewise, the single-item scale for overall quality of life was not considered in our analysis.

Urinary and sexual function items are scored on 3 to 11-point Likert scales. The pre-surgery form of the STAR instrument includes seven urinary function items and the post-surgery form includes five (questions 2 and 4 are common to both). For sexual function, the same six items are included in both pre- and post-surgery forms. Item scores are summed for the urinary and sexual function domains and then transformed to scores ranging from 0 to 100.

We made two wording changes to the STAR instrument. First, our data collection also included the EPIC-26 questionnaire (not reported in the present paper) which overlaps with some STAR items. Where an item existed in both questionnaires, we used the EPIC wording. These minor wording changes are unlikely to substantially change the performance of the item. Second, the standard updated version of STAR has a time frame of six months pre-operatively for both sexual and urinary function, four weeks post operatively for sexual function and one week post operatively for urinary function. To ensure consistency across time for both urinary and sexual function domains, we used a 4-week recall period throughout. We considered this long enough for all problems to be noticed and/or resolved. All items were administered at all time points.

,

Data analysis

Overall fit to the model: For each scale, we evaluated whether the observed responses were significantly different to the responses expected Based on the Rasch model (significant chi square statistic).

Item threshold ordering: For a higher level of functioning on each item, the probability of 'endorsing' a higher response category (on the Likert scale) should increase and the probability of endorsing a lower response category decrease. If each response category in turn (0, 1, 2, 3, 4, 5) has the highest probability of endorsement with increasing levels of functioning, the 'thresholds' between the categories (0-1, 1-2, 2-3, 3-4, 4-5) show a logical order. Thresholds are the location on the scale where the two adjacent response categories have equal probability (50%) of endorsement.

Empirically, however, thresholds can be disordered (e.g. 0-1, 2-3, 1-2), indicating that the response categories do not work as intended. This can be because an item has ambiguous wording or has labels on the response scale that are not sufficiently distinct. We evaluated whether the response categories are working as intended by a visual inspection of the 'category probability curves'.

Item fit validity: The items of the scale must work together ('fit') as a conformable set both clinically and statistically. Clinically, the item ordering along the continuum should make sense and statistically the items need to satisfy specified criteria. Otherwise, it is inappropriate to sum item responses to reach a total score and consider the total score as a measure of the construct. When items do not work together ('misfit') in this way, the validity of a scale needs to be questioned.

We evaluated the fit of each item to the Rasch model by inspecting its 'fit residual' (acceptable range of +/- 2.5) and considering the related Chi-square value. We also assessed visually how closely the observed 'class interval mean scores' follow the expected values in the 'item characteristic curve'. Class intervals are groupings of approximately equal numbers of respondents who have about the same level of functioning.

Differential item functioning (DIF): Stability of the item locations is assessed by evaluating 'differential item functioning (DIF)'. DIF occurs when different groups

within the sample, for example patients of different age, respond differently to an item, despite having the same level of functioning. DIF is identified through an ANOVA main effect for 'person factors', for example age by an interaction between the person factor and the class intervals.

In both the urinary and sexual function scales, we evaluated DIF by age, ethnicity, relationship status and number of co-morbidities. For items that were scored both before and after surgery (two items for the urinary function scale and all six items for the sexual functioning scale), we also evaluated DIF by time point.

Local response dependency: The response to one item should not directly influence the response to another. If 'item response-dependency' happens, measurement estimates can be biased, and reliability, indicated by the 'person separation index', is artificially increased. Local response dependency is evaluated by examining the residual correlations between the items after the Rasch factor they have in common has been partialled out. A correlation coefficient with a value larger than 0.20 above the average of all the item residual correlations indicates potential local response dependency (25).

Reliability (internal consistency): Reliability was examined using the 'person separation index' which is a statistic comparable to the Cronbach's alpha, often used in traditional methods based on CTT. It quantifies how reliably the scale distinguishes between respondents. It is computed from the variation among person locations relative to the standard error of estimate for each individual respondent (16). Higher person separation index values indicate better reliability; a value >0.70 at group level and >0.85 at individual level indicates adequate reliability (20).

Scale to sample targeting: 'Scale-to-sample targeting' describes the match between the range of the construct measured by the items and the range of the construct in the sample of patients. This is evaluated by the 'person-item distribution' which compares the difference between 'person locations' and 'item threshold locations' on the underlying ruler, that captures for example urinary or sexual function. Any gaps in item threshold locations, in particular at the low and high ends of the scale, means

that the functioning of respondents located in that gap area cannot be measured precisely. In other words, their scores will have a relatively large standard error of measurement, because their estimation is severely affected by missing information.

All p-values were adjusted for sample size (n=500) as Chi-square values are sensitive to sample size (26)). As a sensitivity analysis forb the correction of the p-values, we repeated all analyses on a random sub-sample of 400. Furthermore, Bonferroni corrections for multiple testing were also applied. All analyses were carried using RUMM 2030 (26).

Patient and Public Involvement Statement

Patients and the public were not involved in the design, conduct or dissemination of the project, except as participants in the study.

RESULTS

Study sample

Overall, 971 men were eligible, of whom 873 were approached, 714 were interested and 431 men completed the online consent form, giving an overall recruitment rate of 44.4%.

Of the 431 patients who provided consent, 403 patients (93.5%) completed at least one valid questionnaire. A total of 366 valid questionnaires were completed at baseline, 222 questionnaires were completed at one month after surgery and 181 questionnaires at three months after surgery. Table 1 describes the characteristics of the 403 patients included in this analysis. These patients had a mean age of 63 years (SD 6.7; range 41 – 78 years), were predominantly white or white-British (79.7%), and were mostly married or living with a partner (76.7%).

Overall fit to the model

The overall Chi-square statistic indicated that neither the urinary function nor the sexual function scale fit the Rasch model (urinary function, chi square=207.04p<0.001; sexual function, chi square=341.98; p<0.001).

Item threshold ordering

Both urinary and sexual function scales had items with disordered thresholds, indicating that the response options were not working as intended. The urinary function scale had disordered thresholds for 7 of the 10 items. For these 7 disordered items, the category probability plots in Figures 1a-1g illustrate that this is mainly a problem with the middle response options, suggesting that the wording was not clear or that the difference between categories was not well understood. For example, for Q3 of the urinary function scale ('Over the last 4 weeks, how often have you found you stopped and started again several times when you urinated?') there is no point at which threshold 2 ('About half the time') and threshold 3 ('Less than half the time') are the most likely to occur. If the response options were working as intended, the probability of each threshold should come in order.

All six of the sexual function items are disordered. This means that none of the response scales are working as they were intended. Figures 2a-2f indicate that it is mainly thresholds 2 and 3 that are disordered, suggesting that the middle categories of the response scales are not well understood and may need to be re-worded.

Item fit validity

Both the urinary and sexual function scales contained items that did not fit the model, when considering together their fit residual, Chi-square value, and the item characteristic curve (fit residuals and chi Square values for all items are reported in Table 2). One urinary function item (Q23) failed all three criteria indicating misfit to the model. Two further items failed one or two criteria (Q3 and Q7) indicating a broader problem with item fit.

Five sexual function items failed all three criteria (Table 2) and the remaining item failed one of the three criteria suggesting further problems with item fit.

Differential item functioning (DIF)

Overall, items in both scales were stable (invariant) across different groups for age, ethnicity, relationship status and number of co-morbidities. However, both scales contained items that were unstable across time, with the sexual function scale containing a greater number of unstable items.

One urinary item (Q23) showed DIF across time points (p<0.001). Patients' response to this item were systematically higher at 3 months post-op compared to 1 month post-surgery, despite having equal underlying levels of urinary function.

Five sexual function items (Q9, Q10, Q11, Q12, Q13) showed DIF by time (p<0.001)

Local response dependency

Both scales contained pairs of items that were dependent on each other, but the sexual function scale showed greater local dependency. Four pairs of urinary function items showed local dependency: Q3 (stopped and started again) and Q4 (difficulty postponing urination) (residual correlation = 0.10); Q5 (weak urinary stream) and Q6 (push or strain to begin urination) (residual correlation = 0.04); Q19 (leaking urine) and Q21 (number of pads per day) (residual correlation = 0.32); Q21 (number of pads per day) and Q23 (urinary problem overall) (residual correlation = 0.13).

Four pairs of sexual function items showed local dependency with relatively high residual correlations: Q10 (erection during sexual activity) and Q11 (erections hard enough for penetration) (residual correlation = 0.30), Q12 (able to penetrate) and Q13 (maintain erection after penetration) (residual correlation = 0.59), Q12 (able to penetrate) and Q14 (maintain erection to completion) (residual correlation = 0.55), Q13 (maintain erection after penetration) and Q14 (maintain erection to completion) (residual correlation = 0.51).

Reliability

Internal consistency was acceptable at group level for both scales (urinary function scale: person separation index = 0.75; sexual function scale: person separation index = 0.82).

Scale-to-sample targeting

The person-item distribution of the urinary function scale was relatively poor, though better than the targeting for the sexual function scale (Figure 3a). Although the middle of the person distribution is reasonably well matched by items, both extremes of the distribution have few items. This means that for men located at the lower end of the scale (including many men at one month after surgery) and at the higher end of the scale (including many men before surgery) the level of functioning cannot be precisely measured.

The targeting for the sexual functioning scale was also poor (Figure 3b). In particular, most items are located in the centre of the scale whereas the distribution of people is quite wide. This means that the sexual function for men located at the higher end of the scale (often men before surgery) and the lower end of the scale (most of the men after surgery) is very imprecisely measured.

The sensitivity analyses conducted on a random sub-sample (n=400) broadly showed a pattern of results that was comparable with the whole sample results presented here. The targeting diagrams, disordered thresholds, pattern of local response dependency and DIF are very similar. The pattern of item fit is slightly improved in the random sub-sample and as expected (because the n is smaller) fewer items meet the criteria for mis-fit based on fit residuals and the significance of the chi square. However, the pattern of variation across items for mis-fit is in the same direction as the original sample.

DISCUSSION

Our analyses have identified that neither the urinary function items nor the sexual function items from the STAR instrument can be placed on a common metric that is robust for comparisons before and after surgery. Furthermore, a number of anomalies have been identified that suggest the scales are not working as intended. There is an inadequate match between location of items and the distribution of the patients, suggesting that the underlying constructs that the scales purport to measure are not clear. Consequently, the items do not measure the men's function very accurately. The response categories for many items are not consistently used, some items do not work with the others as a conformable set and some items are not stable over time.

These results indicate that in its current form the items in the STAR instrument do not provide an adequate ruler to monitor urinary or sexual function in clinical practice. These problems are likely to make the estimation of an individual patient's outcome after surgery less accurate and precise and using the questionnaire in its current form therefore carries a risk of misrepresenting actual urinary and sexual outcomes.

Our results demonstrate that the risk of inaccurate estimation of outcomes using STAR is likely to be most pronounced for men with either very good or very poor outcomes. The poor scale-to-sample targeting, particularly for the sexual functioning scale, also means that this problem is exacerbated for men with better function before surgery and worse function after surgery, creating clear problems for the interpretation of change scores that are supposed to capture the impact of surgery. Further, both scales have items that showed DIF by time providing further evidence that it is not meaningful to compare scores before and after surgery or compare scores taken at different times after surgery.

In the short term, some of the identified deficiencies can be addressed using post-hoc statistical techniques to re-score the disordered thresholds (16, 20) or to resolve for the uniform DIF (24) and local response dependency (20). However, a more robust solution would be to conduct qualitative research with men who have experienced radical prostatectomy to understand why the questions are not well understood and why the response options are not used in the way that was intended. Qualitative research should also explore which areas of content are missing and how items could be formulated to address these gaps. A revised version based on these findings would then need to be psychometrically evaluated again to determine how well the amendments to content and scoring have addressed the identified problems.

This study is the first to use robust modern psychometric methods such as Rasch analysis to determine the measurement properties of a prostate cancer-specific PROM (21) and to evaluate its suitability to collect PROMs for use in clinical practice at the level of individual patients. It has allowed us to scrutinise each aspect of the questionnaire and to identify carefully which aspects work well and which do not.

In our study, the questionnaire was completed at home rather than in clinic and there may be differences between our setting and the setting that was originally used to developed the instrument, especially with respect to the amount of support men received whilst completing the questionnaire.

We also used a different time frame and did not adapt the questions to UK English (as we wanted to evaluate the original questionnaire in its US wording). Yet, it is likely that the anomalies identified in relation to item misfit and inconsistent threshold ordering reflected ambiguous and confusing wording rather than simply linguistic differences between US and UK English.

All of our analyses used a Bonferroni correction to adjust the p-values. Although widely used, this approach has been criticised as conservative. This may therefore have had the effect of under-estimating the number of anomalies found in these two scales.

CONCLUSION

Using Rasch analysis as a diagnostic tool, we have identified several shortcomings of the STAR instrument. In their current form both the urinary function and the sexual function scales have issues that need to be resolved before STAR can be used with confidence in clinical practice. The sexual function scale in particular is unlikely to provide precise estimates for the outcomes experienced by men after radical prostatectomy. For both scales, the underlying construct is not clear and needs further investigation.

Our results demonstrate the need to evaluate the suitability of any PROMs in routine clinical practice, including for example the EPIC-26 that is currently being implemented in prostate cancer care in the UK (10, 11), using modern psychometric methods to identify and address deficiencies that affect their psychometric performance.

Without appropriate psychometric scrutiny and related further development where needed, the use of PROMs in routine clinical practice may significantly misrepresent the true clinical outcomes for patients. PROMs that produce inaccurate and imprecise scores have limited value for clinicians who aim to respond to the needs of their patients. Inaccurate and imprecise scores will also undermine the guiding role that PROMs can have for patients who want to contribute to the management of their own condition. Without progress in development in this area we lose the opportunity to demonstrate the benefit of new technology. This will be detrimental to patients both now and in the future.

Acknowledgements

This work is part of the TrueNTH UK Post Surgery project. TrueNTH is a global initiative by the Movember Foundation, which aims to identify and demonstrate the best and most cost-effective models for improving prostate cancer survivorship care and support. We also thank Dr Jolijn Hendriks for helpful comments on a draft of the manuscript

Author contributorship

EP wrote the first draft of the paper and SS and EP were responsible for the psychometric analysis. CM and JvdM were responsible for the design of the study. All authors contributed to drafting the manuscript and have approved the final version.

Competing Interests Statement

The authors have no conflicts of interest relevant to this article to disclose.

Ethics Statement

Ethical approval for the study was obtained (Study Title: True NTH UK – Post Surgical Follow-up; REC Reference 15/SC/0451).

Funding Statement

This work is funded by Prostate Cancer UK. The funder had no role in any of the following: design and conduct of the study, data collection and management, data analysis and interpretation, or preparation, approval and review of the manuscript.

Financial Disclosure

The authors have no financial relationships relevant to this article to disclose.

Data Availability Statement

No additional data available

Figure captions:

Figures 1a-1g: Urinary Function Category Probability Curves for disordered items
Figures 2a-2f: Sexual Function Category Probability Curves for disordered items
Figure 3a: Urinary Function Person-Item Distribution (targeting)
Figure 3b: Sexual Function Person-Item Distribution (targeting)

References

1. Black N. Patient reported outcome measures could help transform healthcare. BMJ : British Medical Journal. 2013;346.

2. England N. Patient Reported Outcome Measures (PROMs) 2017 [cited 2017 Oct 2017]. Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/proms/.

3. Wales N. Patient Reported Outcome Measures 2017 [cited 2017 Oct 2017]. Available from: https://proms.nhs.wales/.

4. Chard J, Kuczawski M, Black N, van der Meulen J. Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. BMJ. 2011;343.

5. Browne J, Jamieson L, Lewsey J, van der Meulen J, Black N, Cairns J, et al. Patient reported outcome measures (PROMs) in elective surgery. Report to the Department of Health. 2007;12.

6. Baumhauer JF, Bozic KJ. Value-based Healthcare: Patient-reported Outcomes in Clinical Decision Making. Clinical Orthopaedics and Related Research®. 2016;474(6):1375-8.

7. Jason B. Liu M, Andrea L. Pusic, MD, MHS, FACS, Larissa K. Temple, MD, MSc, FACS and Clifford Y. Ko, MD, MS, MSHS, FACS, FASCRS. Patient-reported outcomes in surgery: Listening to patients improves quality of care: Bulletin of the American College of Surgeons; 2017 [Oct 2017]. Available from: http://bulletin.facs.org/2017/03/patient-reported-outcomes-in-surgery-listening-to-patients-improves-quality-of-care/.

8. Vickers AJ, Savage CJ, Shouery M, Eastham JA, Scardino PT, Basch EM. Validation study of a web-based assessment of functional recovery after radical prostatectomy. Health and Quality of Life Outcomes. 2010;8:82.

9. Brundage MD, Barbera L, McCallum F, Howell DM. A pilot evaluation of the expanded prostate cancer index composite for clinical practice (EPIC-CP) tool in Ontario. Qual Life Res. 2018 Oct 31. doi: 10.1007/s11136-018-2034-x. [Epub ahead of print] PubMed PMID: 30382479.

10. Madaan S, Reekhaye A, McFarlane J. Survivorship and prostate cancer: the TrueNTH supported self-management programme. Trends in Urology & Men's Health, January/February 2016:21-24. https://cdn.movember.com/uploads/files/Our%20Work/truenth-supported-self-management-programme-movember-foundation.pdf

11. TrueNTH, a Movember initiative https://prostatecanceruk.org/for-health-professionals/our-projects/truenth

12. US Food and Drug Administration. Guidance for industry on patient-reported outcome measures: Use in medicinal product development to support labeling claims. 2009.

13. Chassany O, Sagnier P, Marquis P, Fullerton S, Aaronson N, Group ERIoQoLA. Patient-reported outcomes: the example of health-related quality of life—a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. Drug Information Journal. 2002;36(1):209-38.

14. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research. 2002;11(3):193-205.

15. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures

used in patient-centered outcomes and comparative effectiveness research. Qual Life Res. 2013;22(8):1889-905.

16. 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):200.

17. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. The Lancet Neurology. 2007;6(12):1094-105.

18. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Press M, editor: MESA Press; 1960.

19. Wright BD, G. M. Rating scale analysis: Rasch measurement. Chicago: MESA; 1982.

20. 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;57(8):1358-62.

21. Protopapa E, van der Meulen J, Moore CM, Smith SC. Patient-reported outcome (PRO) questionnaires for men who have radical surgery for prostate cancer: a conceptual review of existing instruments. BJU international. 2017;120(4):468-81.

22. Wright B. Rack and stack: time 1 vs. time 2. Rasch measurement transactions. 2003;17(1):905-6.

23. Andrich, D., & Marais, I. (2019). A Course in Rasch Measurement Theory. https://doi.org/10.1007/978-981-13-7496-8

24. Andrich D, Luo G, BE. S. Interpreting RUMM2020. Perth, WA: RUMM Laboratory2004.

25. Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q 3 : Identification of Local Dependence in the Rasch Model Using Residual Correlations. Applied Psychological Measurement, 41(3), 178–194. https://doi.org/10.1177/0146621616677520

26. Andrich D, Sheridan B. RUMM2030. Perth, WA: RUMM Laboratory Pty Ltd; 1997-2017.

Table 1: Sample characteristics of the 403 patients who completed at least one valid questionnaire

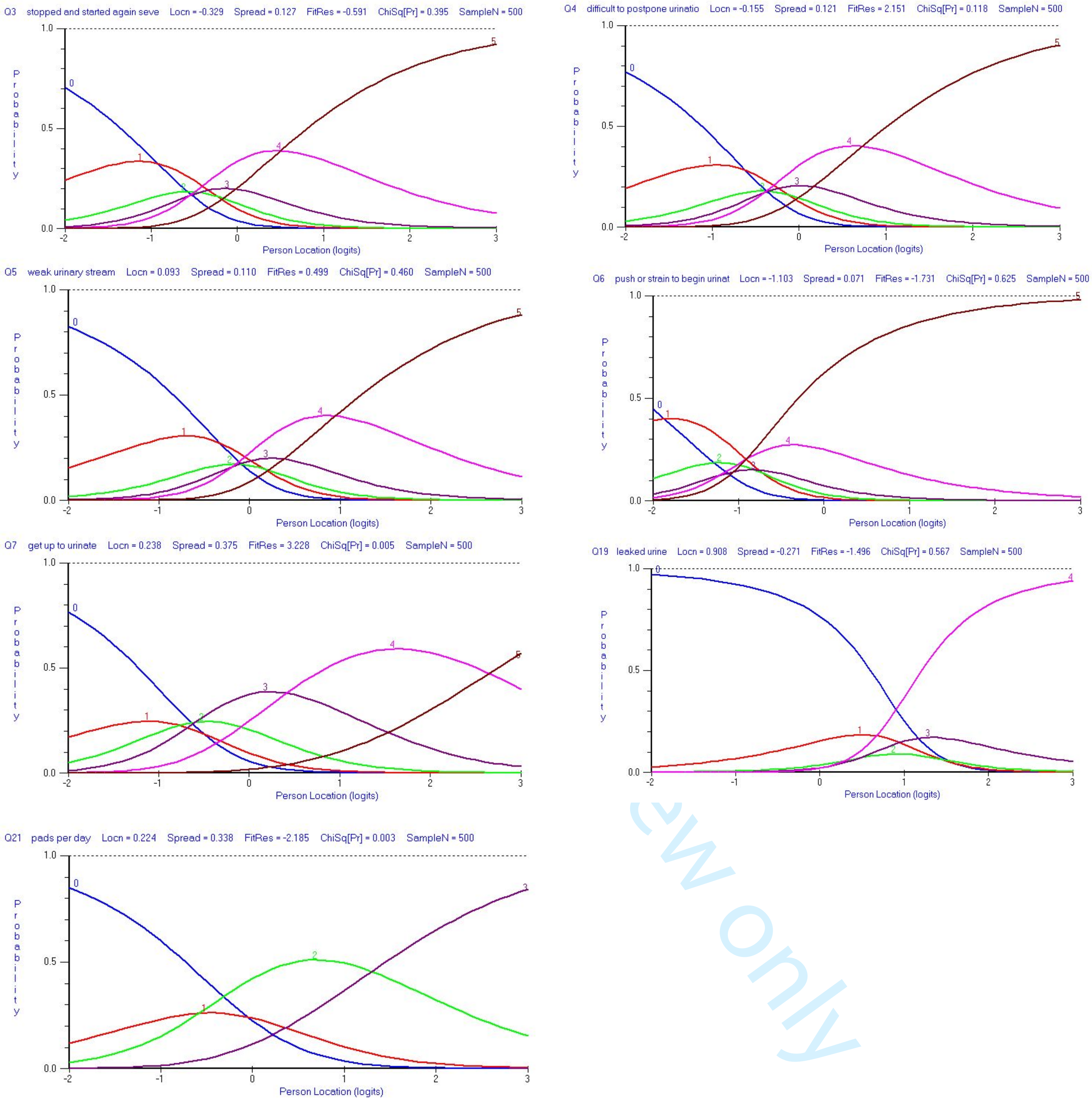
Sample characteristics			N (%)
Age			
<60			123 (30.5)
60-66			131 (32.5)
>66			149 (37.0)
Ethnicity			
White/White British			321 (79.6)
Other ethnicity			45 (11.2)
Missing			37 (9.2)
Relationship			
Married or living with a partner			309 (76.7)
Other			55 (13.6)
Missing			39 (9.7)
No. of co-morbidities			
0			133 (33.0)
1			164 (40.7)
>2			69 (17.1)
Missing			39 (9.2)

Table 2: Urinary function & sexual function – item fit

Urinary function Item	Location	SE	FitResid	DF	ChiSq	DF	Prob	ICC
Q1 non-complete emptying	-0.492	0.053	-3.077	294.78	15.691	8	0.047026	
Q2 urinate again less than 2hours	0.33	0.035	0.21	598.61	6.623	9	0.676341	
Q3 stopped & started again	-0.329	0.05	-0.591	293.95	8.401	8	0.39534	
Q4 difficult to postpone	-0.155	0.033	2.151	595.31	14.137	9	0.117529	
Q5 weak stream	0.093	0.045	0.499	293.95	7.733	8	0.460021	
Q6 push /strain to begin	-1.103	0.068	-1.731	294.78	6.196	8	0.625333	
Q7 get up in night to urinate	0.238	0.054	3.228	295.6	22.219	8	0.004526	
Q19 leaked urine	0.908	0.047	-1.496	303.83	7.676	9	0.567147	
Q21 pads per day	0.224	0.062	-2.185	304.66	25.356	9	0.002602	
Q23 urinary function - problem	0.287	0.054	-3.157	300.54	27.97	9	0.000965	Questionable
Sexual function Item	**Location**	**SE**	**FitResid**	**DF**	**ChiSq**	**DF**	**Prob**	**ICC**
Q9 confidence to get & keep erection	0.119	0.056	5.814	400.73	135.786	8	0	Questionable
Q10 erection during sexual activity	-0.496	0.049	-2.28	399.9	30.792	8	0.000153	
Q11 erections hard enough for penetration	-0.266	0.05	-3.729	399.08	48.484	8	0	Questionable
Q12 able to penetrate partner	0.195	0.052	-6.208	397.43	49.952	8	0	Questionable
Q13 maintain erection after penetration	0.32	0.053	-5.078	396.61	41.819	8	0.000001	Questionable
Q14 maintain erection to completion	0.129	0.05	-5.152	398.25	35.149	8	0.000025	Questionable

Highlighted items fail criteria

Figures 1a-1g: Urinary Function Category Probability Curves for disordered items



Figures 2a-2f: Sexual Function Category Probability Curves for disordered items

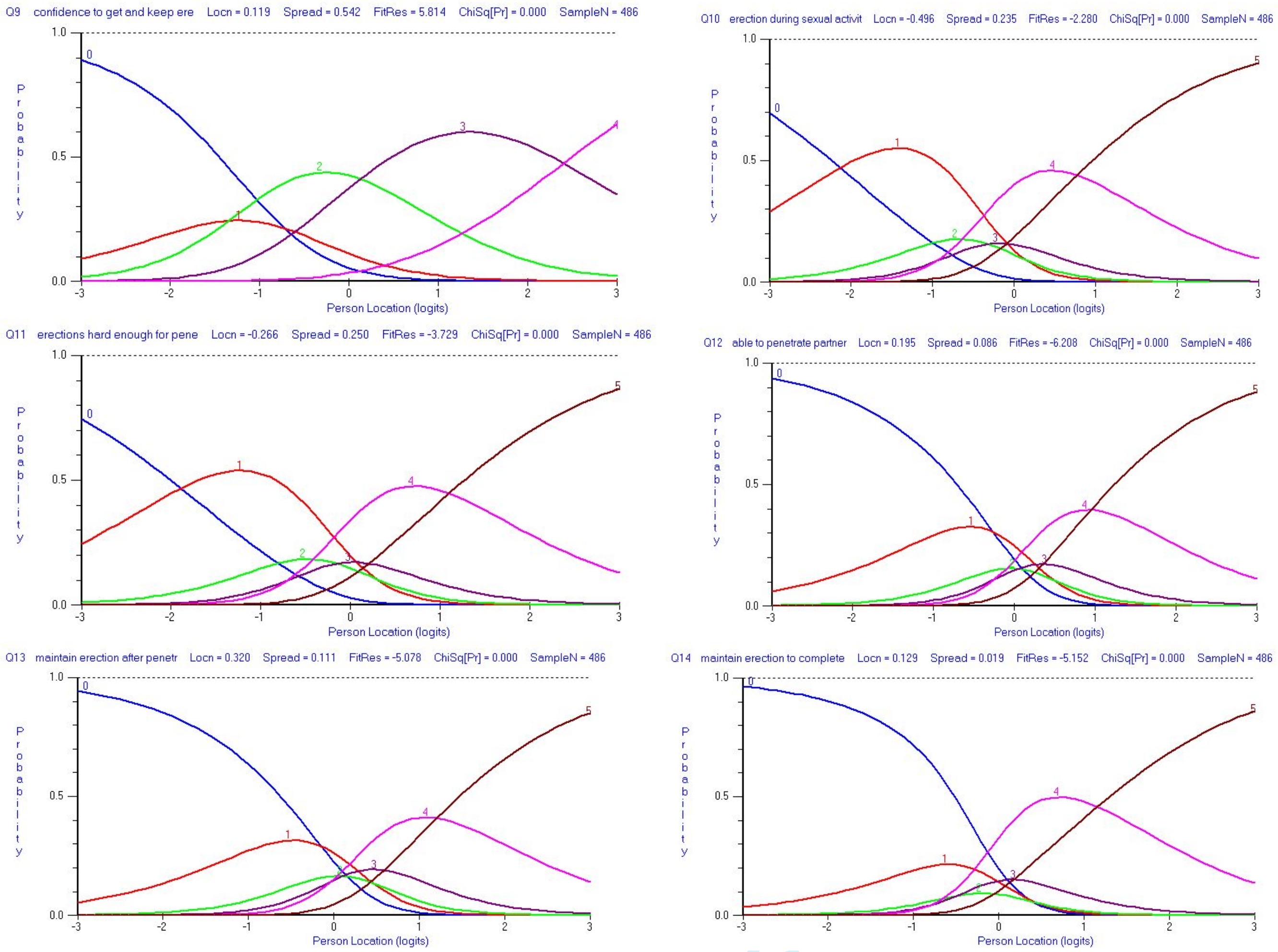


Figure 3a: Urinary Function Person-Item Distribution (targeting)

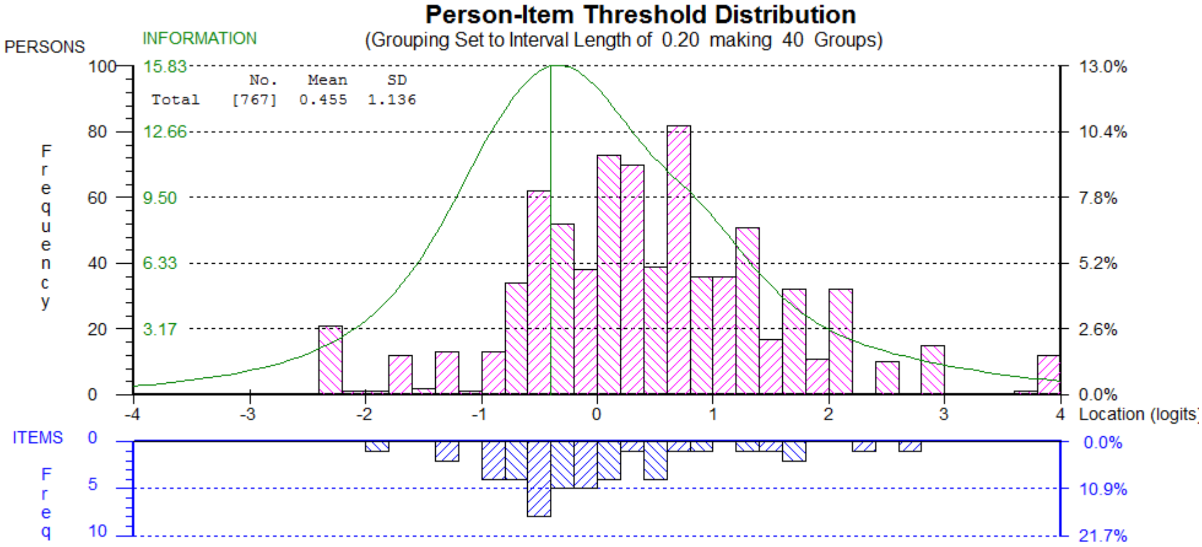
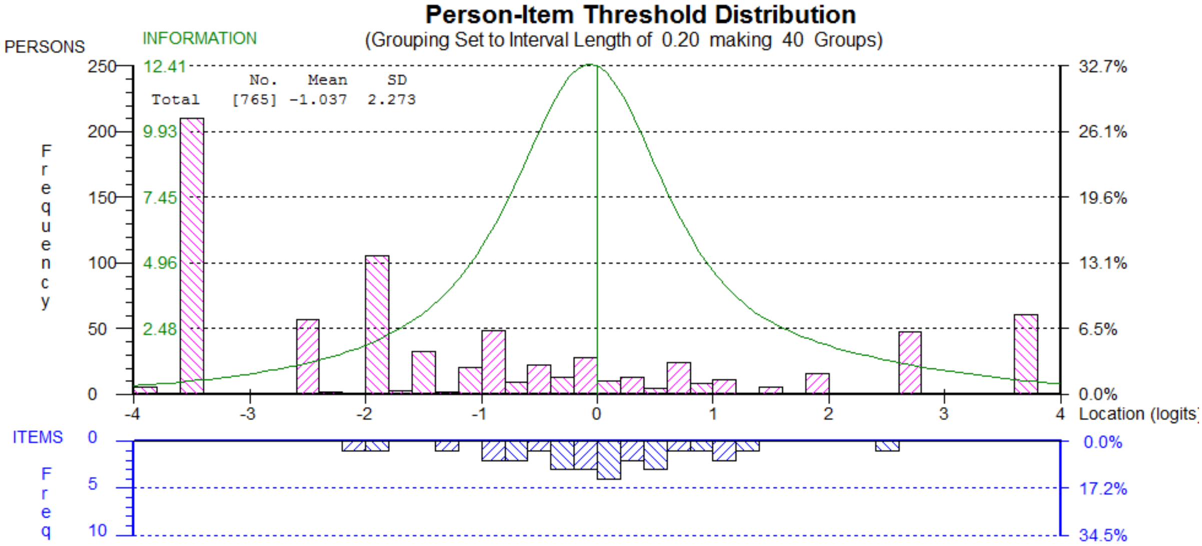


Figure 3b: Sexual Function Person-Item Distribution (targeting)



STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract see page 1 (*b*) Provide in the abstract an informative and balanced summary of what was done and what was found see page 2
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported see page 3-5
Objectives	3	State specific objectives, including any prespecified hypotheses see page 5-6
Methods		
Study design	4	Present key elements of study design early in the paper See page 2 and 6
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection See page 6
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up See page 6 (*b*) For matched studies, give matching criteria and number of exposed and unexposed *N/A*
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable See page 6-7
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group See page 6-7
Bias	9	Describe any efforts to address potential sources of bias See page 11 and 12
Study size	10	Explain how the study size was arrived at N/A
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why N/A
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding (*b*) Describe any methods used to examine subgroups and interactions (*c*) Explain how missing data were addressed (*d*) If applicable, explain how loss to follow-up was addressed (*e*) Describe any sensitivity analyses See page 10
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study,

		completing follow-up, and analysed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram See pages 10-11
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarise follow-up time (eg, average and total amount) See page 11
Outcome data	15*	Report numbers of outcome events or summary measures over time See page 6
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period N/A
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses See pages 10-13
Discussion		
Key results	18	Summarise key results with reference to study objectives See page 13
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias See page 3 and 15
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence See pages 13-15
Generalisability	21	Discuss the generalisability (external validity) of the study results N/A
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based See page 16

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

BMJ Open

Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-035436.R3
Article Type:	Original research
Date Submitted by the Author:	01-Jul-2020
Complete List of Authors:	Protopapa, Eva; London School of Hygiene and Tropical Medicine, Department of Health Services Research & Policy van der Meulen, Jan; London School of Hygiene and Tropical Medicine, Department of Health Services Research & Policy Moore, Caroline; University College London, Division of Surgery and Interventional Sciences Smith, Sarah; London School of Hygiene & Tropical Medicine, HSRP
<b>Primary Subject Heading</b>:	Urology
Secondary Subject Heading:	Surgery
Keywords:	UROLOGY, Prostate disease < UROLOGY, SURGERY

Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis

Evangelina Protopapa[1] Jan van der Meulen,[1] Caroline M Moore,[2] Sarah Smith[1]

1 London School of Hygiene and Tropical Medicine, London UK

2 University College London, London, UK

Author email addresses:

e.protopapa@ucl.ac.uk

JanvanderMeulen@lshtm.ac.uk

caroline.moore@ucl.ac.uk

Sarah.Smith@lshtm.ac.uk

Address for correspondence;

Dr Sarah Smith, Associate Professor in Psychology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine 15-17 Tavistock Place, London WC1H 9SH. Tel: 0207 9272038

Email: sarah.smith@lshtm.ac.uk

Word count:

Abstract: 297

Text: 4165

Keywords:
PROM, psychometric, Rasch, prostate, surgery

ABSTRACT

OBJECTIVES: To evaluate the psychometric properties (and identify specific anomalies to be resolved) of urinary and sexual function scales of the STAR instrument for use in clinical practice with individual men using Rasch analysis.

DESIGN: Prospective cohort study

SETTING: 9 UK surgery centres in secondary care

PARTICIPANTS: 403 men diagnosed with prostate cancer and completed at least one questionnaire immediately before and at 1 or 3 months after a radical prostatectomy.

PRIMARY AND SECONDARY OUTCOMES: STAR instrument before surgery and 1 and 3 months afterwards.

RESULTS: Neither scale fitted the Rasch model (both scales p<0.001). Both urinary (7 items) and sexual function (6 items) had disordered thresholds, suggesting response categories are not working as intended. Both scales (3 urinary items; 5 sexual function items) showed problems with item fit (large fit residuals, significant chi square, inspection of item characteristic curves (ICC)). Both scales showed items that were unstable over time (DIF by time). Both scales (4 pairs of items in each scale) showed local response dependency (residual correlations >0.2 above the average). Internal consistency was acceptable at the group level for both scales. Targeting was poor for both scales, indicating an inadequate match between location of items and the distribution of the patients, suggesting that the underlying constructs that the scales purport to measure are not clear.

CONCLUSION: Using Rasch analysis as a diagnostic tool, we identified that both the urinary and the sexual function scales have issues that need to be resolved before STAR can be used with confidence in clinical practice. The sexual function scale in particular is unlikely to provide precise estimates for the outcomes experienced by men after radical prostatectomy. These results demonstrate the need to evaluate the suitability of any PROM before implementation in routine clinical practice, preferably using modern psychometric methods**.**

STRENGTHS AND LIMITATIONS OF THIS STUDY

- used modern psychometric methods (based on Rasch measurement Theory) to determine if it is appropriate to use a total function score to describe a patient's sexual or urinary function.
- determined how well the items in each scale reflect the experience of men who report the questionnaire
- determined specific anomalies in the scores that suggest that the scales are not being used and understood in the way that was intended
- did not change the items in the questionnaire based on our findings and so did not evaluate any potential improvement such changes would make

INTRODUCTION

The use of patient-reported outcome measures (PROMs) has rapidly increased (1-3). In the UK, PROMs are routinely collected for several areas of elective surgery to evaluate the outcomes in *groups of patients*, receiving a particular treatment or treated in a specific hospital (4, 5). Similar approaches are under consideration for other conditions.

However, there is a lack of evidence about the extent to which clinicians can use PROMs to make their clinical practice more responsive to *individual patients'* needs. Also, it has been suggested that PROMs can play an important role for patients as they can help to inform ways in which patients can self-manage their condition (6, 7).

A web-based tool known as STAR ('Symptom Tracking and Reporting') (8) has been developed at the Memorial Sloan-Kettering Cancer Center (New York, US) to monitor outcomes of radical prostate cancer treatment in individual patients. This

instrument is used to inform both surgeons and men about functional outcomes after surgery, such as urinary, sexual and bowel function improvement or deterioration. Its development is just one example of the implementation of PROMs in prostate cancer practice to inform both clinicians and patients (9-11).

The STAR instrument was not designed to compare men's functional status before and after surgery because different questions are included in the pre- and post-treatment STAR questionnaires. This means that the assessment before surgery is on a different 'ruler' compared to after surgery and therefore there is no clear way of understanding what the change means. However in practice, for example in the English national PROMs programme, pre- and post-treatment PROMs are often compared to monitor the impact of elective surgery (2).

Instruments such as STAR aim to measure specific 'constructs'. It is important these instruments have adequate psychometric properties, otherwise they may produce scores that are 'inaccurate' (prone to systematic error) or 'imprecise' (prone to random error), making it difficult to understand what the observed scores mean and even more difficult to interpret changes over time.

The criteria that must be met to ensure that PROMs are robust are well established (12-15). They ensure that the 'scale' that results from adding up responses to individual questions ('items') relates to a clear underlying construct, as distinct from descriptive responses or simple counts of how many times a symptom occurs.

Like most health-related PROMs, the STAR instrument has been developed using traditional psychometric methods based on classical test theory (CTT). There are important limitations to these methods (16). First, the scales developed using CTT produce 'ordinal scores', where the difference between two adjacent scores at different points on the scale may not be equal. This poses a problem because most statistical analyses assume scores have interval properties where differences between adjacent scores are equal across the entire scale. When scales are based on ordinal scores, changes over time are especially difficult to interpret. Second, the scores can only be interpreted for groups of patients, because measures of statistical uncertainty of these scores (e.g. 'standard errors') are only computed at group level,

which limits their use for individual patients (17). Third, the performance of scales is dependent on the particular sample in which they are used. This makes it difficult to compare studies and, even more importantly, undermines further the interpretation of changes over time.

Modern psychometric methods, such as those based on ‘item response theory’ (IRT) or ‘Rasch measurement theory’, provide a way of overcoming these challenges. Both are mathematical modelling approaches transforming ordinal scales into interval measures, provided that certain model-related criteria are met. But whereas IRT takes a statistical approach of adding parameters to the model in order to improve its fit to the data, the Rasch paradigm takes a theory-driven approach that investigates why the data do not fit the Rasch model (18-20). The Rasch paradigm, however, keeps central the conceptual underpinning of the instrument and provides a clear set of diagnostic statistics that can help to identify anomalies in its scores.

Instruments developed using these modern psychometric methods have four main advantages over CTT-based instruments. First, they have the potential to generate truly interval scores, thus improving the accuracy and precision with which change over time can be evaluated. Second, measures of statistical uncertainty can be estimated for scores of individual respondents, meaning that the interpretation of scores at patient level is more meaningful. Third, it is possible to produce scales that do not depend on a particular sample’s characteristics. Fourth, they can create a model that contains both pre-and post-surgery items, and therefore all items can be calibrated on the same ruler. The usual pre and post-treatment scores can still be derived but calibrated in such a way that they can be properly compared.

In a systematic review of seven prostate cancer-specific PROMs, including the STAR instrument (21), we identified that modern psychometric methods had not been used to evaluate the psychometric properties of these instruments. In this study, we therefore used Rasch analysis to estimate urinary and sexual function for individual men based on responses to the STAR instrument that were provided by men immediately before and up to three months after radical prostate cancer surgery. In so doing, we aimed to identify anomalies that should be addressed to make the STAR instrument, or any other PROM that aims to monitor changes in outcomes over time after prostate cancer surgery, suitable for use in routine clinical practice.

We performed analyses based on Rasch measurement theory to determine if it is appropriate to use a total function score to describe a patient's sexual or urinary function. As comparisons are often made between pre- and post-surgery scores, we aimed to determine if the seven pre-surgery and five post-surgery items could be placed on the same measurement ruler. If they can, then meaningful comparisons can be made across time. To do this, we 'stacked' the data, in other words, we added the baseline and follow-up scores for each patient as separate records (22).

The analyses aimed to answer a number of questions. First, has a measurement ruler been successfully constructed? Second, have the people been successfully measured? Third, is the scale-to-sample targeting adequate? The approach to each of these questions is explained briefly below. A more extensive explanation of Rasch measurement theory can be found in a recent overviews (23).

METHODS

Setting and participants

Participants were recruited between November 2015 and March 2017 from nine centres that perform radical prostatectomy by any method (open, laparoscopic-assisted or robotic-assisted) in the UK. Men were eligible if they were diagnosed with prostate cancer, scheduled to have a radical prostatectomy, and had sufficient English language to understand the information about the study and complete the required online questionnaire.

The clinical team at each centre identified and approached eligible patients, informed them about the study, and registered those who were interested in taking part on the secure online portal. Registered patients received their login details by text or email and logged on to the portal to complete the consent form. Once patients had consented, they were directed to the online questionnaire. Patients were invited to

complete the questionnaire before surgery, and at one, three, six and 12 months after surgery.

Instrument

The STAR instrument consists of four domains: sexual function, urinary function, bowel function, and overall quality of life. Our analysis focused on the urinary and sexual function scales obtained immediately before and one and three months after surgery. We excluded the bowel scale from psychometric analyses as with only two items it had insufficient content to be considered a scale. Likewise, the single-item scale for overall quality of life was not considered in our analysis.

Urinary and sexual function items are scored on 3 to 11-point Likert scales. The pre-surgery form of the STAR instrument includes seven urinary function items and the post-surgery form includes five (questions 2 and 4 are common to both). For sexual function, the same six items are included in both pre- and post-surgery forms. Item scores are summed for the urinary and sexual function domains and then transformed to scores ranging from 0 to 100.

We made two wording changes to the STAR instrument. First, our data collection also included the EPIC-26 questionnaire (not reported in the present paper) which overlaps with some STAR items. Where an item existed in both questionnaires, we used the EPIC wording. These minor wording changes are unlikely to substantially change the performance of the item. Second, the standard updated version of STAR has a time frame of six months pre-operatively for both sexual and urinary function, four weeks post operatively for sexual function and one week post operatively for urinary function. To ensure consistency across time for both urinary and sexual function domains, we used a 4-week recall period throughout. We considered this long enough for all problems to be noticed and/or resolved. All items were administered at all time points.

,

Data analysis

Overall fit to the model: For each scale, we evaluated whether the observed responses were significantly different to the responses expected Based on the Rasch model (significant chi square statistic).

Item threshold ordering: For a higher level of functioning on each item, the probability of 'endorsing' a higher response category (on the Likert scale) should increase and the probability of endorsing a lower response category decrease. If each response category in turn (0, 1, 2, 3, 4, 5) has the highest probability of endorsement with increasing levels of functioning, the 'thresholds' between the categories (0-1, 1-2, 2-3, 3-4, 4-5) show a logical order. Thresholds are the location on the scale where the two adjacent response categories have equal probability (50%) of endorsement.

Empirically, however, thresholds can be disordered (e.g. 0-1, 2-3, 1-2), indicating that the response categories do not work as intended. This can be because an item has ambiguous wording or has labels on the response scale that are not sufficiently distinct. We evaluated whether the response categories are working as intended by a visual inspection of the 'category probability curves'.

Item fit validity: The items of the scale must work together ('fit') as a conformable set both clinically and statistically. Clinically, the item ordering along the continuum should make sense and statistically the items need to satisfy specified criteria. Otherwise, it is inappropriate to sum item responses to reach a total score and consider the total score as a measure of the construct. When items do not work together ('misfit') in this way, the validity of a scale needs to be questioned.

We evaluated the fit of each item to the Rasch model by inspecting its 'fit residual' (acceptable range of +/- 2.5) and considering the related Chi-square value. We also assessed visually how closely the observed 'class interval mean scores' follow the expected values in the 'item characteristic curve'. Class intervals are groupings of approximately equal numbers of respondents who have about the same level of functioning.

Differential item functioning (DIF): Stability of the item locations is assessed by evaluating 'differential item functioning (DIF)'. DIF occurs when different groups

within the sample, for example patients of different age, respond differently to an item, despite having the same level of functioning. DIF is identified through an ANOVA main effect for 'person factors', for example age by an interaction between the person factor and the class intervals.

In both the urinary and sexual function scales, we evaluated DIF by age, ethnicity, relationship status and number of co-morbidities. For items that were scored both before and after surgery (two items for the urinary function scale and all six items for the sexual functioning scale), we also evaluated DIF by time point.

Local response dependency: The response to one item should not directly influence the response to another. If 'item response-dependency' happens, measurement estimates can be biased, and reliability, indicated by the 'person separation index', is artificially increased. Local response dependency is evaluated by examining the residual correlations between the items after the Rasch factor they have in common has been partialled out. A correlation coefficient with a value larger than 0.20 above the average of all the item residual correlations indicates potential local response dependency (24).

Reliability (internal consistency): Reliability was examined using the 'person separation index' which is a statistic comparable to the Cronbach's alpha, often used in traditional methods based on CTT. It quantifies how reliably the scale distinguishes between respondents. It is computed from the variation among person locations relative to the standard error of estimate for each individual respondent (16). Higher person separation index values indicate better reliability; a value >0.70 at group level and >0.85 at individual level indicates adequate reliability (20).

Scale to sample targeting: 'Scale-to-sample targeting' describes the match between the range of the construct measured by the items and the range of the construct in the sample of patients. This is evaluated by the 'person-item distribution' which compares the difference between 'person locations' and 'item threshold locations' on the underlying ruler, that captures for example urinary or sexual function. Any gaps in item threshold locations, in particular at the low and high ends of the scale, means

that the functioning of respondents located in that gap area cannot be measured precisely. In other words, their scores will have a relatively large standard error of measurement, because their estimation is severely affected by missing information.

All p-values were adjusted for sample size (n=500) as Chi-square values are sensitive to sample size (25)). As a sensitivity analysis for the correction of the p-values, we repeated all analyses on a random sub-sample of 400. Furthermore, Bonferroni corrections for multiple testing were also applied. All analyses were carried using RUMM 2030 (26).

Patient and Public Involvement Statement

Patients and the public were not involved in the design, conduct or dissemination of the project, except as participants in the study.

RESULTS

Study sample

Overall, 971 men were eligible, of whom 873 were approached, 714 were interested and 431 men completed the online consent form, giving an overall recruitment rate of 44.4%.

Of the 431 patients who provided consent, 403 patients (93.5%) completed at least one valid questionnaire. A total of 366 valid questionnaires were completed at baseline, 222 questionnaires were completed at one month after surgery and 181 questionnaires at three months after surgery. Table 1 describes the characteristics of the 403 patients included in this analysis. These patients had a mean age of 63 years (SD 6.7; range 41 – 78 years), were predominantly white or white-British (79.7%), and were mostly married or living with a partner (76.7%).

Overall fit to the model

The overall Chi-square statistic indicated that neither the urinary function nor the sexual function scale fit the Rasch model (urinary function, chi square=207.04p<0.001; sexual function, chi square=341.98; p<0.001).

Item threshold ordering

Both urinary and sexual function scales had items with disordered thresholds, indicating that the response options were not working as intended. The urinary function scale had disordered thresholds for 7 of the 10 items. For these 7 disordered items, the category probability plots in Figures 1a-1g illustrate that this is mainly a problem with the middle response options, suggesting that the wording was not clear or that the difference between categories was not well understood. For example, for Q3 of the urinary function scale ('Over the last 4 weeks, how often have you found you stopped and started again several times when you urinated?') there is no point at which threshold 2 ('About half the time') and threshold 3 ('Less than half the time') are the most likely to occur. If the response options were working as intended, the probability of each threshold should come in order.

All six of the sexual function items are disordered. This means that none of the response scales are working as they were intended. Figures 2a-2f indicate that it is mainly thresholds 2 and 3 that are disordered, suggesting that the middle categories of the response scales are not well understood and may need to be re-worded.

Item fit validity

Both the urinary and sexual function scales contained items that did not fit the model, when considering together their fit residual, Chi-square value, and the item characteristic curve (fit residuals and chi Square values for all items are reported in Table 2). One urinary function item (Q23) failed all three criteria indicating misfit to the model. Two further items failed one or two criteria (Q3 and Q7) indicating a broader problem with item fit.

Five sexual function items failed all three criteria (Table 2) and the remaining item failed one of the three criteria suggesting further problems with item fit.

Differential item functioning (DIF)

Overall, items in both scales were stable (invariant) across different groups for age, ethnicity, relationship status and number of co-morbidities. However, both scales contained items that were unstable across time, with the sexual function scale containing a greater number of unstable items.

One urinary item (Q23) showed DIF across time points (p<0.001). Patients' response to this item were systematically higher at 3 months post-op compared to 1 month post-surgery, despite having equal underlying levels of urinary function.

Five sexual function items (Q9, Q10, Q11, Q12, Q13) showed DIF by time (p<0.001)

Local response dependency

Both scales contained pairs of items that were dependent on each other, but the sexual function scale showed greater local dependency. Four pairs of urinary function items showed local dependency: Q3 (stopped and started again) and Q4 (difficulty postponing urination) (residual correlation = 0.10); Q5 (weak urinary stream) and Q6 (push or strain to begin urination) (residual correlation = 0.04); Q19 (leaking urine) and Q21 (number of pads per day) (residual correlation = 0.32); Q21 (number of pads per day) and Q23 (urinary problem overall) (residual correlation = 0.13).

Four pairs of sexual function items showed local dependency with relatively high residual correlations: Q10 (erection during sexual activity) and Q11 (erections hard enough for penetration) (residual correlation = 0.30), Q12 (able to penetrate) and Q13 (maintain erection after penetration) (residual correlation = 0.59), Q12 (able to penetrate) and Q14 (maintain erection to completion) (residual correlation = 0.55), Q13 (maintain erection after penetration) and Q14 (maintain erection to completion) (residual correlation = 0.51).

Reliability

Internal consistency was acceptable at group level for both scales (urinary function scale: person separation index = 0.75; sexual function scale: person separation index = 0.82).

Scale-to-sample targeting

The person-item distribution of the urinary function scale was relatively poor, though better than the targeting for the sexual function scale (Figure 3a). Although the middle of the person distribution is reasonably well matched by items, both extremes of the distribution have few items. This means that for men located at the lower end of the scale (including many men at one month after surgery) and at the higher end of the scale (including many men before surgery) the level of functioning cannot be precisely measured.

The targeting for the sexual functioning scale was also poor (Figure 3b). In particular, most items are located in the centre of the scale whereas the distribution of people is quite wide. This means that the sexual function for men located at the higher end of the scale (often men before surgery) and the lower end of the scale (most of the men after surgery) is very imprecisely measured.

The sensitivity analyses conducted on a random sub-sample (n=400) broadly showed a pattern of results that was comparable with the whole sample results presented here. The targeting diagrams, disordered thresholds, pattern of local response dependency and DIF are very similar. The pattern of item fit is slightly improved in the random sub-sample and as expected (because the n is smaller) fewer items meet the criteria for mis-fit based on fit residuals and the significance of the chi square. However, the pattern of variation across items for mis-fit is in the same direction as the original sample.

DISCUSSION

Our analyses have identified that neither the urinary function items nor the sexual function items from the STAR instrument can be placed on a common metric that is robust for comparisons before and after surgery. Furthermore, a number of anomalies have been identified that suggest the scales are not working as intended. There is an inadequate match between location of items and the distribution of the patients, suggesting that the underlying constructs that the scales purport to measure are not clear. Consequently, the items do not measure the men's function very accurately. The response categories for many items are not consistently used, some items do not work with the others as a conformable set and some items are not stable over time.

These results indicate that in its current form the items in the STAR instrument do not provide an adequate ruler to monitor urinary or sexual function in clinical practice. These problems are likely to make the estimation of an individual patient's outcome after surgery less accurate and precise and using the questionnaire in its current form therefore carries a risk of misrepresenting actual urinary and sexual outcomes.

Our results demonstrate that the risk of inaccurate estimation of outcomes using STAR is likely to be most pronounced for men with either very good or very poor outcomes. The poor scale-to-sample targeting, particularly for the sexual functioning scale, also means that this problem is exacerbated for men with better function before surgery and worse function after surgery, creating clear problems for the interpretation of change scores that are supposed to capture the impact of surgery. Further, both scales have items that showed DIF by time providing further evidence that it is not meaningful to compare scores before and after surgery or compare scores taken at different times after surgery.

In the short term, some of the identified deficiencies can be addressed using post-hoc statistical techniques to re-score the disordered thresholds (16, 20) or to resolve for the uniform DIF (25) and local response dependency (20). However, a more robust solution would be to conduct qualitative research with men who have experienced radical prostatectomy to understand why the questions are not well understood and why the response options are not used in the way that was intended. Qualitative research should also explore which areas of content are missing and how items could be formulated to address these gaps. A revised version based on these findings would then need to be psychometrically evaluated again to determine how well the amendments to content and scoring have addressed the identified problems.

This study is the first to use robust modern psychometric methods such as Rasch analysis to determine the measurement properties of a prostate cancer-specific PROM (21) and to evaluate its suitability to collect PROMs for use in clinical practice at the level of individual patients. It has allowed us to scrutinise each aspect of the questionnaire and to identify carefully which aspects work well and which do not.

In our study, the questionnaire was completed at home rather than in clinic and there may be differences between our setting and the setting that was originally used to developed the instrument, especially with respect to the amount of support men received whilst completing the questionnaire.

We also used a different time frame and did not adapt the questions to UK English (as we wanted to evaluate the original questionnaire in its US wording). Yet, it is likely that the anomalies identified in relation to item misfit and inconsistent threshold ordering reflected ambiguous and confusing wording rather than simply linguistic differences between US and UK English.

All of our analyses used a Bonferroni correction to adjust the p-values. Although widely used, this approach has been criticised as conservative. This may therefore have had the effect of under-estimating the number of anomalies found in these two scales.

CONCLUSION

Using Rasch analysis as a diagnostic tool, we have identified several shortcomings of the STAR instrument. In their current form both the urinary function and the sexual function scales have issues that need to be resolved before STAR can be used with confidence in clinical practice. The sexual function scale in particular is unlikely to provide precise estimates for the outcomes experienced by men after radical prostatectomy. For both scales, the underlying construct is not clear and needs further investigation.

Our results demonstrate the need to evaluate the suitability of any PROMs in routine clinical practice, including for example the EPIC-26 that is currently being implemented in prostate cancer care in the UK (10, 11), using modern psychometric methods to identify and address deficiencies that affect their psychometric performance.

Without appropriate psychometric scrutiny and related further development where needed, the use of PROMs in routine clinical practice may significantly misrepresent the true clinical outcomes for patients. PROMs that produce inaccurate and imprecise scores have limited value for clinicians who aim to respond to the needs of their patients. Inaccurate and imprecise scores will also undermine the guiding role that PROMs can have for patients who want to contribute to the management of their own condition. Without progress in development in this area we lose the opportunity to demonstrate the benefit of new technology. This will be detrimental to patients both now and in the future.

Acknowledgements

This work is part of the TrueNTH UK Post Surgery project. TrueNTH is a global initiative by the Movember Foundation, which aims to identify and demonstrate the best and most cost-effective models for improving prostate cancer survivorship care and support. We also thank Dr Jolijn Hendriks for helpful comments on a draft of the manuscript

Author contributorship

EP wrote the first draft of the paper and SS and EP were responsible for the psychometric analysis. CM and JvdM were responsible for the design of the study. All authors contributed to drafting the manuscript and have approved the final version.

Competing Interests Statement

The authors have no conflicts of interest relevant to this article to disclose.

Ethics Statement

Ethical approval for the study was obtained (Study Title: True NTH UK – Post Surgical Follow-up; REC Reference 15/SC/0451).

Funding Statement

This work is funded by Prostate Cancer UK. The funder had no role in any of the following: design and conduct of the study, data collection and management, data analysis and interpretation, or preparation, approval and review of the manuscript.

Financial Disclosure

The authors have no financial relationships relevant to this article to disclose.

Data Availability Statement

No additional data available

Figure captions:

Figures 1a-1g: Urinary Function Category Probability Curves for disordered items
Figures 2a-2f: Sexual Function Category Probability Curves for disordered items
Figure 3a: Urinary Function Person-Item Distribution (targeting)
Figure 3b: Sexual Function Person-Item Distribution (targeting)

References

1. Black N. Patient reported outcome measures could help transform healthcare. BMJ : British Medical Journal. 2013;346.

2. England N. Patient Reported Outcome Measures (PROMs) 2017 [cited 2017 Oct 2017]. Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/proms/.

3. Wales N. Patient Reported Outcome Measures 2017 [cited 2017 Oct 2017]. Available from: https://proms.nhs.wales/.

4. Chard J, Kuczawski M, Black N, van der Meulen J. Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. BMJ. 2011;343.

5. Browne J, Jamieson L, Lewsey J, van der Meulen J, Black N, Cairns J, et al. Patient reported outcome measures (PROMs) in elective surgery. Report to the Department of Health. 2007;12.

6. Baumhauer JF, Bozic KJ. Value-based Healthcare: Patient-reported Outcomes in Clinical Decision Making. Clinical Orthopaedics and Related Research®. 2016;474(6):1375-8.

7. Jason B. Liu M, Andrea L. Pusic, MD, MHS, FACS, Larissa K. Temple, MD, MSc, FACS and Clifford Y. Ko, MD, MS, MSHS, FACS, FASCRS. Patient-reported outcomes in surgery: Listening to patients improves quality of care: Bulletin of the American College of Surgeons; 2017 [Oct 2017]. Available from: http://bulletin.facs.org/2017/03/patient-reported-outcomes-in-surgery-listening-to-patients-improves-quality-of-care/.

8. Vickers AJ, Savage CJ, Shouery M, Eastham JA, Scardino PT, Basch EM. Validation study of a web-based assessment of functional recovery after radical prostatectomy. Health and Quality of Life Outcomes. 2010;8:82.

9. Brundage MD, Barbera L, McCallum F, Howell DM. A pilot evaluation of the expanded prostate cancer index composite for clinical practice (EPIC-CP) tool in Ontario. Qual Life Res. 2018 Oct 31. doi: 10.1007/s11136-018-2034-x. [Epub ahead of print] PubMed PMID: 30382479.

10. Madaan S, Reekhaye A, McFarlane J. Survivorship and prostate cancer: the TrueNTH supported self-management programme. Trends in Urology & Men's Health, January/February 2016:21-24. https://cdn.movember.com/uploads/files/Our%20Work/truenth-supported-self-management-programme-movember-foundation.pdf

11. TrueNTH, a Movember initiative https://prostatecanceruk.org/for-health-professionals/our-projects/truenth

12. US Food and Drug Administration. Guidance for industry on patient-reported outcome measures: Use in medicinal product development to support labeling claims. 2009.

13. Chassany O, Sagnier P, Marquis P, Fullerton S, Aaronson N, Group ERIoQoLA. Patient-reported outcomes: the example of health-related quality of life—a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. Drug Information Journal. 2002;36(1):209-38.

14. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research. 2002;11(3):193-205.

15. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures

used in patient-centered outcomes and comparative effectiveness research. Qual Life Res. 2013;22(8):1889-905.

16. 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):200.

17. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. The Lancet Neurology. 2007;6(12):1094-105.

18. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Press M, editor: MESA Press; 1960.

19. Wright BD, G. M. Rating scale analysis: Rasch measurement. Chicago: MESA; 1982.

20. 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;57(8):1358-62.

21. Protopapa E, van der Meulen J, Moore CM, Smith SC. Patient-reported outcome (PRO) questionnaires for men who have radical surgery for prostate cancer: a conceptual review of existing instruments. BJU international. 2017;120(4):468-81.

22. Wright B. Rack and stack: time 1 vs. time 2. Rasch measurement transactions. 2003;17(1):905-6.

23. Andrich, D., & Marais, I. (2019). A Course in Rasch Measurement Theory. https://doi.org/10.1007/978-981-13-7496-8

24. Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q 3 : Identification of Local Dependence in the Rasch Model Using Residual Correlations. Applied Psychological Measurement, 41(3), 178–194. https://doi.org/10.1177/0146621616677520

25. Andrich D, Luo G, BE. S. Interpreting RUMM2020. Perth, WA: RUMM Laboratory2004.

26. Andrich D, Sheridan B. RUMM2030. Perth, WA: RUMM Laboratory Pty Ltd; 1997-2017.

Table 1: Sample characteristics of the 403 patients who completed at least one valid questionnaire

Sample characteristics		N (%)
Age		
<60		123 (30.5)
60-66		131 (32.5)
>66		149 (37.0)
Ethnicity		
White/White British		321 (79.6)
Other ethnicity		45 (11.2)
Missing		37 (9.2)
Relationship		
Married or living with a partner		309 (76.7)
Other		55 (13.6)
Missing		39 (9.7)
No. of co-morbidities		
0		133 (33.0)
1		164 (40.7)
>2		69 (17.1)
Missing		39 (9.2)

Table 2: Urinary function & sexual function – item fit

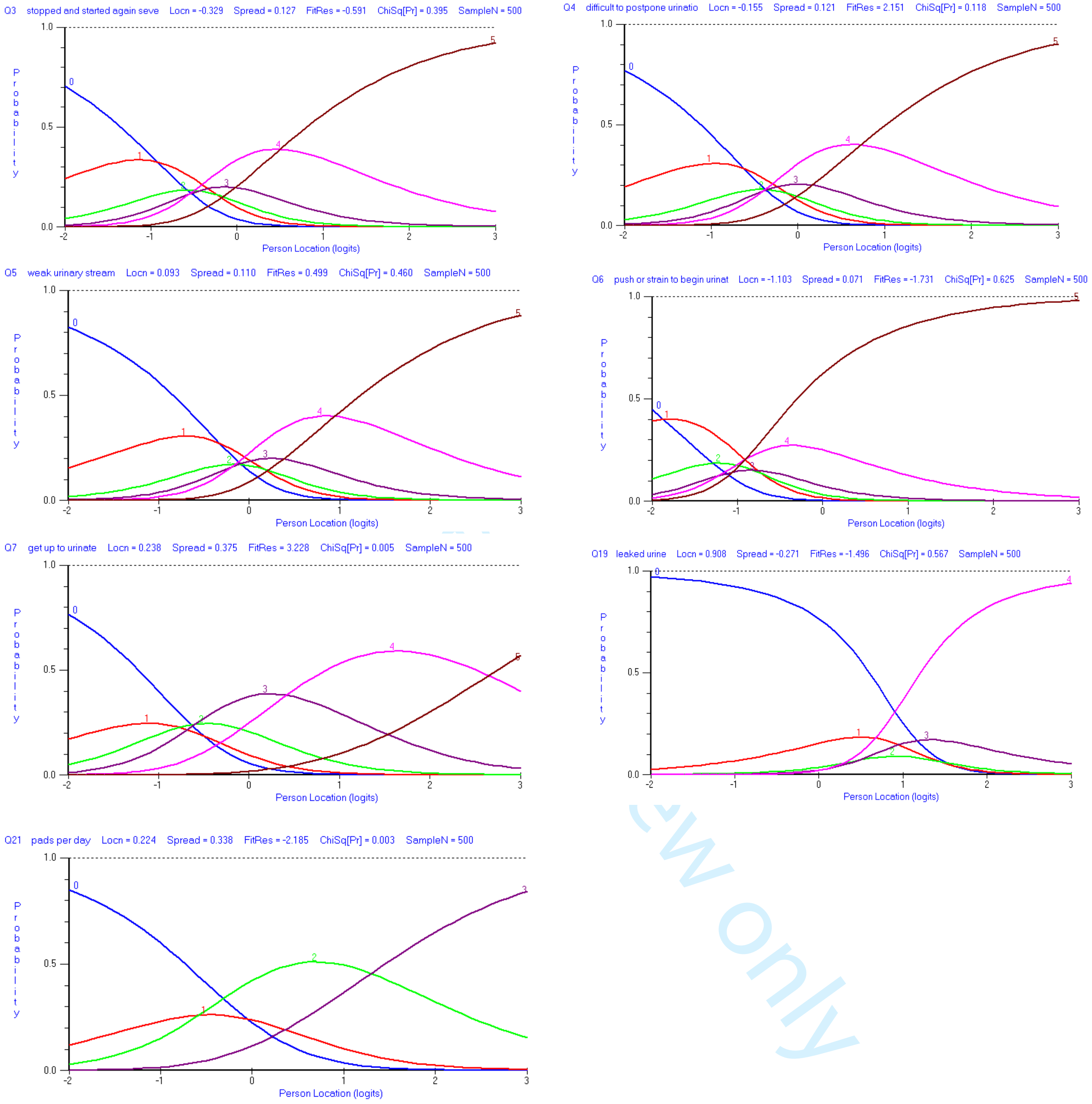
Urinary function Item	Location	SE	Fit Residual	DF	ChiSq	DF	Prob	ICC**
Q1 non-complete emptying	-0.492	0.053	-3.077	294.78	15.691	8	0.047026	
Q2 urinate again less than 2hours	0.33	0.035	0.21	598.61	6.623	9	0.676341	
Q3 stopped & started again	-0.329	0.05	-0.591	293.95	8.401	8	0.395534	
Q4 difficult to postpone	-0.155	0.033	2.151	595.31	14.137	9	0.117529	
Q5 weak stream	0.093	0.045	0.499	293.95	7.733	8	0.460021	
Q6 push /strain to begin	-1.103	0.068	-1.731	294.78	6.196	8	0.625333	
Q7 get up in night to urinate	0.238	0.054	3.228	295.6	22.219	8	0.004526	
Q19 leaked urine	0.908	0.047	-1.496	303.83	7.676	9	0.567147	
Q21 pads per day	0.224	0.062	-2.185	304.66	25.356	9	0.002602	
Q23 urinary function - problem	0.287	0.054	-3.157	300.54	27.97	9	0.000965	Questionable

Sexual function Item	Location	SE	Fit Residual	DF	ChiSq	DF	Prob	ICC**
Q9 confidence to get & keep erection	0.119	0.056	5.814	400.73	135.786	8	0	Questionable
Q10 erection during sexual activity	-0.496	0.049	-2.28	399.9	30.792	8	0.000153	
Q11 erections hard enough for penetration	-0.266	0.05	-3.729	399.08	48.484	8	0	Questionable
Q12 able to penetrate partner	0.195	0.052	-6.208	397.43	49.952	8	0	Questionable
Q13 maintain erection after penetration	0.32	0.053	-5.078	396.61	41.819	8	0.000001	Questionable
Q14 maintain erection to completion	0.129	0.05	-5.152	398.25	35.149	8	0.000025	Questionable

**ICC = Item Characteristic Curve

Highlighted items fail criteria

Figures 1a-1g: Urinary Function Category Probability Curves for disordered items



Figures 2a-2f: Sexual Function Category Probability Curves for disordered items

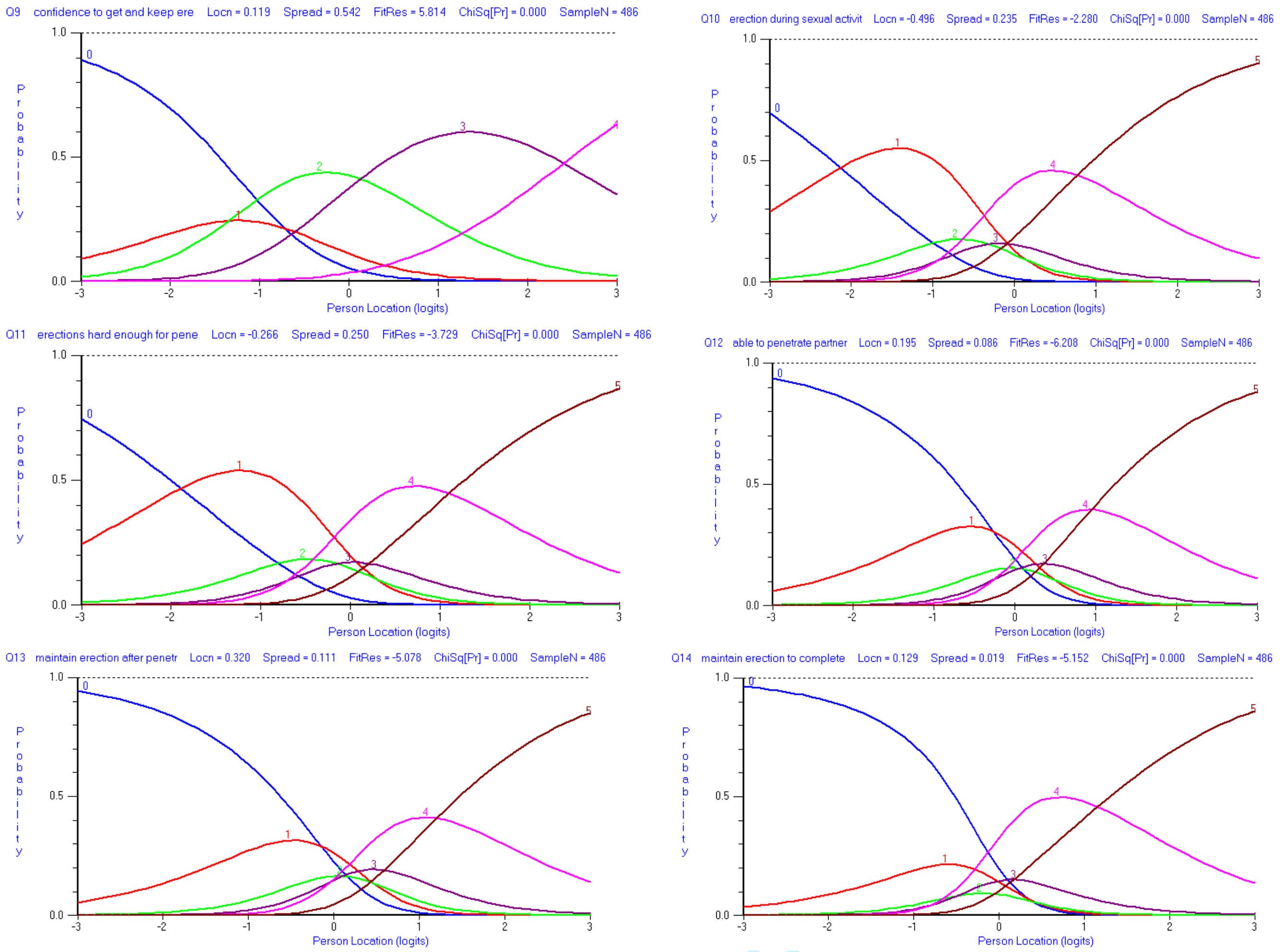


Figure 3a: Urinary Function Person-Item Distribution (targeting)

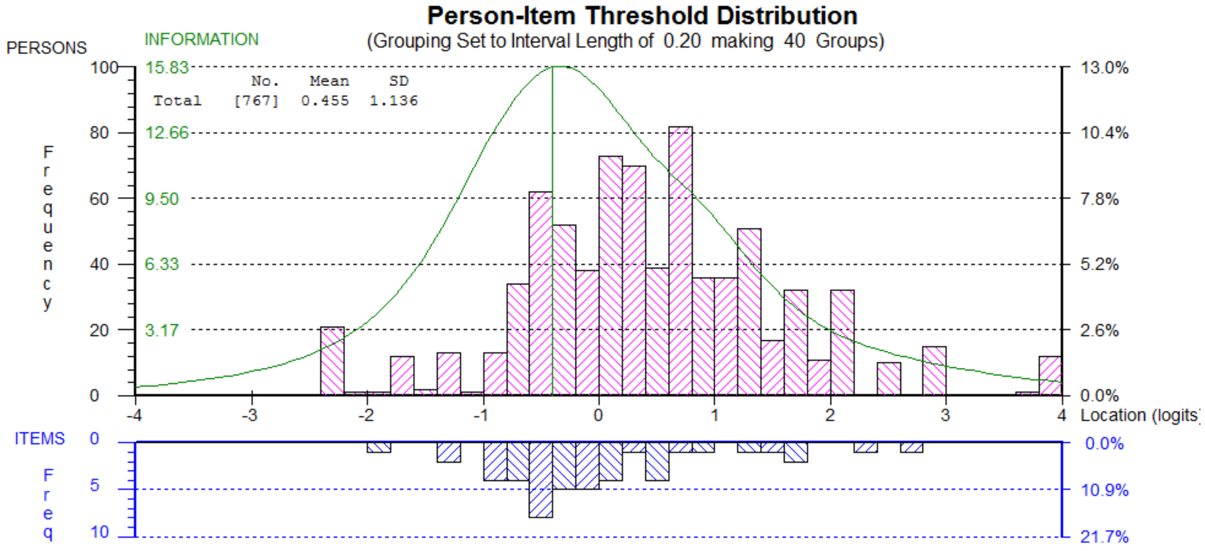
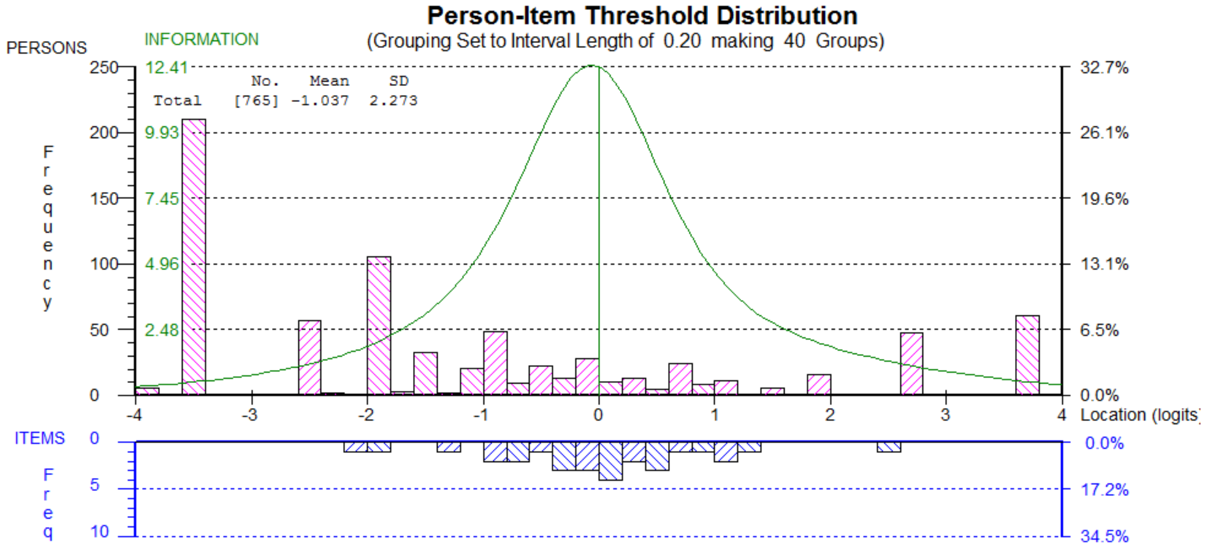


Figure 3b: Sexual Function Person-Item Distribution (targeting)



STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract see page 1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found see page 2
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported see page 3-5
Objectives	3	State specific objectives, including any prespecified hypotheses see page 5-6
Methods		
Study design	4	Present key elements of study design early in the paper See page 2 and 6
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection See page 6
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up See page 6
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed *N/A*
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable See page 6-7
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group See page 6-7
Bias	9	Describe any efforts to address potential sources of bias See page 11 and 12
Study size	10	Explain how the study size was arrived at N/A
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why N/A
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) If applicable, explain how loss to follow-up was addressed
		(*e*) Describe any sensitivity analyses See page 10
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study,

		completing follow-up, and analysed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram See pages 10-11
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarise follow-up time (eg, average and total amount) See page 11
Outcome data	15*	Report numbers of outcome events or summary measures over time See page 6
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period N/A
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses See pages 10-13
Discussion		
Key results	18	Summarise key results with reference to study objectives See page 13
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias See page 3 and 15
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence See pages 13-15
Generalisability	21	Discuss the generalisability (external validity) of the study results N/A
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based See page 16

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.