

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

Eva Protopapa,<sup>1</sup> Jan van der Meulen,<sup>1</sup> Caroline M Moore,<sup>2</sup> Sarah C Smith <sup>1</sup>

**To cite:** Protopapa E, van der Meulen J, Moore CM, *et al*. Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: a Rasch analysis. *BMJ Open* 2020;**10**:e035436. doi:10.1136/bmjopen-2019-035436

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-035436>).

Received 31 October 2019  
Revised 01 July 2020  
Accepted 05 August 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

<sup>1</sup>Department of Health Services Research & Policy, London School of Hygiene & Tropical Medicine, London, UK

<sup>2</sup>Division of Surgery and Interventional Sciences, University College London, London, UK

## Correspondence to

Dr Sarah C Smith;  
Sarah.Smith@lshtm.ac.uk

## ABSTRACT

**Objectives** To evaluate the psychometric properties (and identify specific anomalies to be resolved) of urinary and sexual function scales of the Symptom Tracking and Reporting (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 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 (seven items) and sexual function (six items) had disordered thresholds, suggesting response categories are not working as intended. Both scales (three urinary items; five sexual function items) showed problems with item fit (large fit residuals, significant  $\chi^2$ , an inspection of item characteristic curves). Both scales showed items that were unstable over time (differential item functioning (DIF) by time). Both scales (four 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 the 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 patient-reported outcome measure before implementation in routine clinical practice, preferably using modern psychometric methods.

## INTRODUCTION

The use of patient-reported outcome measures (PROMs) has rapidly increased.<sup>1–3</sup> In the UK, PROMs are routinely collected for several areas of elective surgery to evaluate the outcomes in *groups of patients*, receiving a

## 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.

particular treatment or treated in a specific hospital.<sup>4 5</sup> 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.<sup>6 7</sup>

A web-based tool known as Symptom Tracking and Reporting (STAR)<sup>8</sup> has been developed at the Memorial Sloan-Kettering Cancer Center (New York, USA) 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.<sup>9–11</sup>

The STAR instrument was not designed to compare men's functional status before and after surgery because different questions are included in the pre-treatment and



post-treatment STAR questionnaires. This means that the assessment before surgery is on a different ‘ruler’ compared with 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-treatment and post-treatment PROMs are often compared to monitor the impact of elective surgery.<sup>2</sup>

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.<sup>12–15</sup> 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.<sup>16</sup> 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 (eg, ‘SEs’) are only computed at the group level, which limits their use for individual patients.<sup>17</sup> 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.<sup>18–20</sup> The Rasch paradigm 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 the 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 presurgery and postsurgery items, and therefore all items can be calibrated on the same ruler. The usual pretreatment 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,<sup>21</sup> 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 3 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-surgery 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.<sup>22</sup>

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 recent overviews.<sup>23</sup>

## 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 1, 3, 6 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 1 and 3 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-point 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 presurgery and postsurgery 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 Expanded Prostate Cancer Index Composite (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 6 months preoperatively for both sexual and urinary function, 4 weeks postoperatively for sexual function and 1 week postoperatively 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 from the responses expected based on the Rasch model (significant  $\chi^2$  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 (eg, 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 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  $\pm 2.5$ ) and considering the related  $\chi^2$  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

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 analysis of variance (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 comorbidities. 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.<sup>24</sup>

### 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 SE of estimate for each individual respondent.<sup>16</sup> Higher person separation index values indicate better reliability; a value >0.70 at group level and >0.85 at individual level indicates adequate reliability.<sup>20</sup>

### 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 SE of measurement, because their estimation is severely affected by missing information.

All p values were adjusted for sample size (n=500) as  $\chi^2$  values are sensitive to sample size.<sup>25</sup> As a sensitivity analysis for the correction of the p values, we repeated all analyses on a random subsample of 400. Furthermore, Bonferroni corrections for multiple testing were also applied. All analyses were carried using RUMM 2030.<sup>26</sup>

### 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 1 month after surgery and 181 questionnaires at 3 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%).

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

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

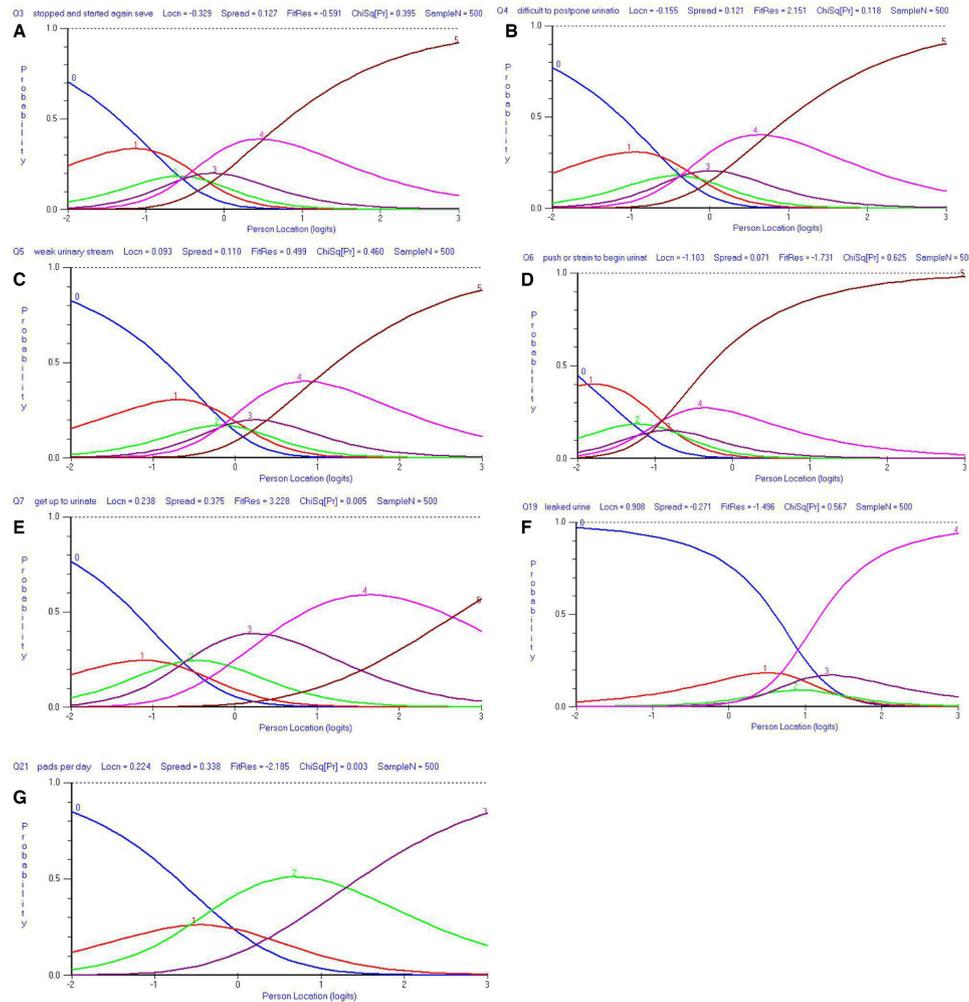
### Overall fit to the model

The overall  $\chi^2$  statistic indicated that neither the urinary function nor the sexual function scale fit the Rasch model (urinary function,  $\chi^2=207.04$ ;  $p<0.001$ ; sexual function,  $\chi^2=341.98$ ;  $p<0.001$ ).

### 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 seven disordered items, the category probability plots in **figure 1A–G** 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. **Figure 2A–F** indicates 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 reworded.



**Figure 1** (A–G) Urinary function category probability curves for disordered items.

### Item fit validity

Both the urinary and sexual function scales contained items that did not fit the model, when considering together their fit residual,  $\chi^2$  value, and the item characteristic curve (fit residuals and  $\chi^2$  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

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 with 1 month postsurgery, 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

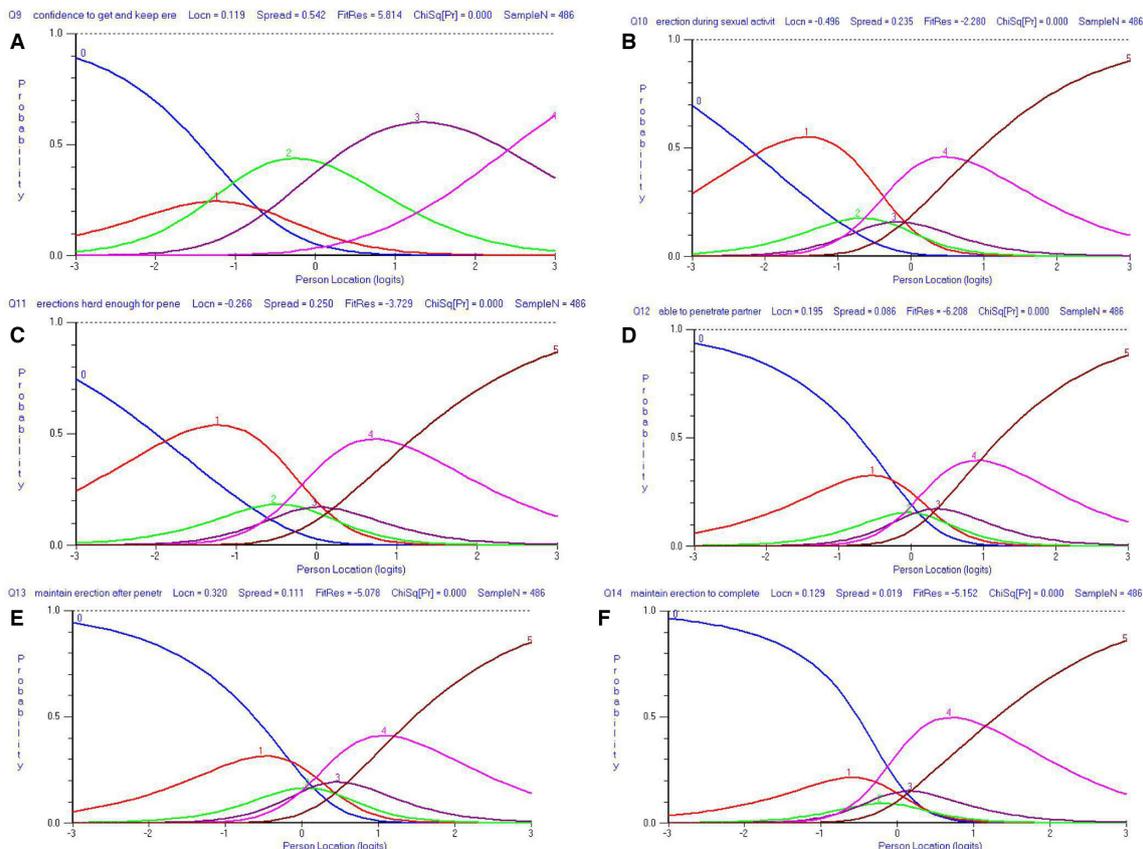
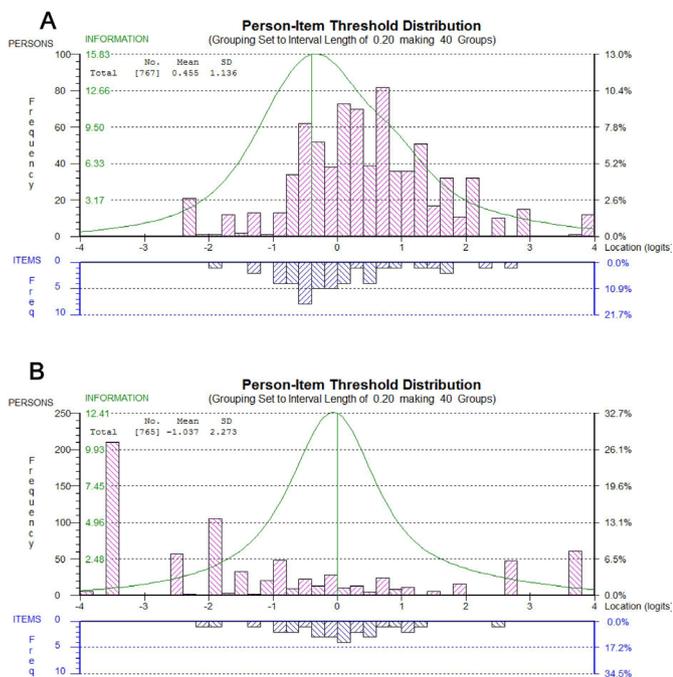


Figure 2 (A–F) Sexual function category probability curves for disordered items.

| Table 2 Urinary function and sexual function—item fit |          |      |              |       |          |    |          |              |
|---|----------|------|--------------|-------|----------|----|----------|--------------|
|   | Location | SE   | Fit residual | DF    | $\chi^2$ | DF | Prob     | ICC          |
| <i>Urinary function item</i>                          |          |      |              |       |          |    |          |              |
| Q1 non-complete emptying                              | -0.492   | 0.05 | -3.077       | 294.8 | 15.691   | 8  | 0.047026 |              |
| Q2 urinate again less than 2 hours                    | 0.33     | 0.04 | 0.21         | 598.6 | 6.623    | 9  | 0.676341 |              |
| Q3 stopped and started again                          | -0.329   | 0.05 | -0.591       | 294   | 8.401    | 8  | 0.39534  |              |
| Q4 difficult to postpone                              | -0.155   | 0.03 | 2.151        | 595.3 | 14.137   | 9  | 0.117529 |              |
| Q5 weak stream  | 0.093    | 0.05 | 0.499        | 294   | 7.733    | 8  | 0.460021 |              |
| Q6 push/strain to begin                               | -1.103   | 0.07 | -1.731       | 294.8 | 6.196    | 8  | 0.625333 |              |
| Q7 get up in night to urinate                         | 0.238    | 0.05 | 3.228        | 295.6 | 22.219   | 8  | 0.004526 |              |
| Q19 leaked urine                                      | 0.908    | 0.05 | -1.496       | 303.8 | 7.676    | 9  | 0.567147 |              |
| Q21 pads per day                                      | 0.224    | 0.06 | -2.185       | 304.7 | 25.356   | 9  | 0.002602 |              |
| Q23 urinary function—problem                          | 0.287    | 0.05 | -3.157       | 300.5 | 27.97    | 9  | 0.000965 | Questionable |
| <i>Sexual function Item</i>                           |          |      |              |       |          |    |          |              |
| Q9 confidence to get and keep erection                | 0.119    | 0.06 | 5.814        | 400.7 | 135.79   | 8  | 0        | Questionable |
| Q10 erection during sexual activity                   | -0.496   | 0.05 | -2.28        | 399.9 | 30.792   | 8  | 0.000153 |              |
| Q11 erections hard enough for penetration             | -0.266   | 0.05 | -3.729       | 399.1 | 48.484   | 8  | 0        | Questionable |
| Q12 able to penetrate partner                         | 0.195    | 0.05 | -6.208       | 397.4 | 49.952   | 8  | 0        | Questionable |
| Q13 maintain erection after penetration               | 0.32     | 0.05 | -5.078       | 396.6 | 41.819   | 8  | 0.000001 | Questionable |
| Q14 maintain erection to completion                   | 0.129    | 0.05 | -5.152       | 398.3 | 35.149   | 8  | 0.000025 | Questionable |

Highlighted items fail criteria.  
ICC, item characteristic curve.



**Figure 3** (A) Urinary function person-Item distribution (targeting). (B) Sexual function person-Item distribution (targeting).

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 1 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 subsample (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

subsample and as expected (because the n is smaller) fewer items meet the criteria for misfit based on fit residuals and the significance of the chi square. However, the pattern of variation across items for misfit 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 the 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 rescore the disordered thresholds<sup>16 20</sup> or to resolve for the uniform DIF<sup>25</sup> and local response dependency.<sup>20</sup> 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<sup>21</sup> 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 develop the instrument, especially with respect to the amount of support men received while 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 the 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,<sup>10 11</sup> 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.

**Twitter** Caroline M Moore @mrsprostate

**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. The authors also thank Dr Jolijn Hendriks for helpful comments on a draft of the manuscript.

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

**Funding** This work is funded by Prostate Cancer UK.

**Disclaimer** 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.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not required.

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

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** No data are available.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

## ORCID iD

Sarah C Smith <http://orcid.org/0000-0002-2013-6963>

## REFERENCES

- Black N. Patient reported outcome measures could help transform healthcare. *BMJ* 2013;346:f167.
- NHS England. Patient reported outcome measures (PROMs), 2017. Available: <https://www.england.nhs.uk/statistics/statistical-work-areas/proms/> [Accessed Oct 2017].
- NHS Wales. Patient reported outcome measures, 2017. Available: <https://proms.nhs.wales/> [Accessed Oct 2017].
- Chard J, Kuczawski M, Black N, et al. Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. *BMJ* 2011;343:d6404.
- Browne J, Jamieson L, Lewsey J, et al. *Patient reported outcome measures (PROMs) in elective surgery*. 12. Report to the Department of Health, 2007.
- Baumhauer JF, Bozic KJ. Value-based healthcare: patient-reported outcomes in clinical decision making. *Clin Orthop Relat Res* 2016;474:1375–8.
- Jason B, Liu M, Andrea L. Patient-reported outcomes in surgery: listening to patients improves quality of care: Bulletin of the American College of surgeons, 2017. Available: <http://bulletin.facs.org/2017/03/patient-reported-outcomes-in-surgery-listening-to-patients-improves-quality-of-care/> [Accessed Oct 2017].
- Vickers AJ, Savage CJ, Shouery M, et al. Validation study of a web-based assessment of functional recovery after radical prostatectomy. *Health Qual Life Outcomes* 2010;8:82.
- Brundage MD, Barbera L, McCallum F, et al. A pilot evaluation of the expanded prostate cancer index composite for clinical practice (EPIC-CP) tool in Ontario. *Qual Life Res* 2019;28:771–82.
- Madaan S, Reekhay A, McFarlane J. Survivorship and prostate cancer: the TrueNTH supported self-management programme. *Trends Urology & Men Health* 2016;7:21–4.
- Prostate Cancer UK. TrueNTH, a Movember initiative. Available: <https://prostatecanceruk.org/for-health-professionals/our-projects/truenth>
- 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, *et al.* 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 Inf J* 2002;36:209–38.
- 14 Aaronson N, Alonso J, Burnam A, *et al.* Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002;11:193–205.
- 15 Reeve BB, Wyrwich KW, Wu AW, *et al.* ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013;22:1889–905.
- 16 Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 2009;13:200.
- 17 Hobart JC, Cano SJ, Zajicek JP, *et al.* Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol* 2007;6:1094–105.
- 18 Rasch G. *Probabilistic models for some intelligence and attainment tests*. MESA Press, 1960.
- 19 Wright BD GM. *Rating scale analysis: Rasch measurement*. Chicago: MESA Press, 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 Rheum* 2007;57:1358–62.
- 21 Protopapa E, van der Meulen J, Moore CM, *et al.* Patient-reported outcome (pro) questionnaires for men who have radical surgery for prostate cancer: a conceptual review of existing instruments. *BJU Int* 2017;120:468–81.
- 22 Wright B. Rack and stack: time 1 vs. time 2. *Rasch Measure Trans* 2003;17:905–6.
- 23 Andrich D, Marais I. *A course in Rasch measurement theory*, 2019.
- 24 Christensen KB, Makransky G, Horton M. Critical values for Yen's  $Q_3$ : identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas* 2017;41:178–94.
- 25 Andrich D, Luo G, BE. S. *Interpreting RUMM2020*. Perth, WA: RUMM Laboratory, 2004.
- 26 Andrich D, Sheridan B. *RUMM2030*. Perth, WA: RUMM Laboratory Pty Ltd, 1997-2017.