Quality of Life Research (2021) 30:2671–2682
https://doi.org/10.1007/s11136-021-02841-1

The development and initial validation of the Breast Cancer Recurrence instrument (BreastCaRe)—a patient-reported outcome measure for detecting symptoms of recurrence after breast cancer

Beverley Lim Høeg[1] · Lena Saltbæk[2,3] · Karl Bang Christensen[4] · Randi Valbjørn Karlsen[1] · Christoffer Johansen[1,5] · Susanne Oksbjerg Dalton[2,3] · Antonia Bennett[6] · Pernille Envold Bidstrup[1,7]

Accepted: 29 March 2021 / Published online: 16 April 2021

Abstract

Purpose Patient-reported outomes (PRO) may facilitate prompt treatment. We describe the development and psychometric properties of the first instrument to monitor for symptoms of breast cancer (BC) recurrence.

Methods This study is nested in the MyHealth randomized trial of nurse-led follow-up based on electronically-collected PROs. We constructed items assessing symptoms of potential recurrence through expert interviews with six BC specialists in Denmark. Semi-structured cognitive interviews were carried out with a patient panel to assess acceptability and comprehensibility. Items were subsequently tested in a population of 1170 women 1–10 years after completing BC treatment. We carried out multiple-groups confirmatory factor analysis (CFA) and Rasch analysis to test dimensionality, local dependence (LD) and differential item functioning (DIF) according to sociodemographic and treatment-related factors. Clinical data was obtained from the Danish Breast Cancer Group registry.

Results Twenty-two items were generated for the Breast Cancer Recurrence instrument (BreastCaRe). Cognitive testing resulted in clearer items. Seven subscales based on general, bone, liver, lung, brain, locoregional and contralateral recurrence symptoms were proposed. Both CFA and Rasch models confirmed the factor structure. No DIF was identified. Five item pairs showed LD but all items were retained to avoid loss of clinical information. Rasch models taking LD into account were used to generate a standardized scoring table for each subscale.

Conclusions The BreastCaRe has good content and structural validity, patient acceptability and measurement invariance. We are preparing to examine the predictive validity of this new instrument.

Keywords Breast cancer follow-up · Multiple groups factor analysis · Patient-reported outcome · Recurrence · Rasch models

✉ Beverley Lim Høeg
bevlim@cancer.dk

[1] Psychological Aspects of Cancer, Danish Cancer Society Research Center, Strandboulevarden 49, 2100 Copenhagen, Denmark

[2] Survivorship and Inequality in Cancer, Danish Cancer Society Research Center, Copenhagen, Denmark

[3] Department of Clinical Oncology & Palliative Care, Zealand University Hospital, Næstved, Denmark

[4] Department of Public Health, University of Copenhagen, Copenhagen, Denmark

[5] Department of Oncology, Copenhagen University Hospital, Copenhagen, Denmark

[6] Cancer Outcomes Research Program, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, USA

[7] Department of Psychology, University of Copenhagen, Copenhagen, Denmark

Plain English summary

Patient-reported outcomes (PRO) are symptoms that are reported directly by the patient, e.g. through a questionnaire, and may help doctors give prompt treatment. Identification of symptoms of recurrence are important during cancer follow-up, but there are currently no PRO questionnaires that assess symptoms of a potential breast cancer recurrence (the cancer returning). We developed a questionnaire of 22 items asking patients about a range of symptoms that may indicate a recurrence. We also interviewed patients to make sure the items were understood as intended. We then analyzed these items using modern psychometric analyses and created the Breast Cancer Recurrence instrument (BreastCaRe) where different groups of items may be used to indicate potential recurrence according to different organ sites. We are currently testing whether the BreastCaRe can be used as a simple cost-effective tool to predict and identify cancer recurrence during cancer follow-up.

Background

Patient-reported outcomes (PRO) are increasingly being used as a clinical tool to improve the treatment and quality-of-life of cancer patients through cost-effective symptom monitoring and prompt intervention [1–3]. This suggests the potential of utilizing PROs not only during treatment, but also in cancer follow-up after completion of treatment. The primary aims of cancer follow-up are the detection of recurrence or new cancers, as well as management of physical and psychosocial late effects [4]. However, although two reviews on PRO measures (PROM) used in cancer clinical practice identified more than 30 measures that were routinely used, these measures either assessed quality of life dimensions or treatment-related effects [5, 6], and none assessed the use of PROs specifically in cancer follow-up.

The utilization of PROs in cancer follow-up may be especially relevant for breast cancer, due to the high survival rates and burgeoning population of breast cancer survivors needing follow-up care [7]. Women with breast cancer have been shown to have a risk ranging from 10% to over 40% of developing a recurrence up to 20 years after completion of treatment [8, 9]. Recurrence can occur as either locoregional recurrence (in the treated breast or lymph nodes), contralateral recurrence (opposite breast), or distant recurrence or metastasis (in other organs) [10, 11]. Previous studies have shown that about 60% of breast cancer recurrences are symptomatic and about 30–40% of recurrences are detected by the woman herself [12–14], suggesting the potential of organizing breast cancer follow-up based on patient-reported symptoms for the detection of possible cancer recurrence.

We identified only one recurrence-related symptom-rating form developed for lung cancer patients. A pilot study ($n = 42$) and subsequently, a randomized trial ($n = 121$), showed earlier detection of disease progression or recurrence and improved survival in patients who filled out the form weekly, when compared to patients who only received planned scans [15, 16]. Although the symptoms and prognosis for lung cancer differ greatly from breast cancer, these results indicate that patient-reported outcomes may potentially be a cost-effective tool for the detection of breast cancer recurrence. However, to our knowledge, there are currently no validated self-reported measures for the detection of symptoms of recurrence in breast cancer survivors.

Depending on the organ site, common symptoms often reported upon suspicion of breast cancer recurrence include lumps or any changes in the feel or appearance of the skin around the breast, pain, dyspnea, weight loss, fatigue, and swelling of lymph nodes [11, 13, 14] which can be operationalized as items in a patient-reported measurement instrument. COSMIN (COnsensus-based Standards for the selection of health status Measurement INstruments) guidelines suggests that the extent to which a new instrument can provide meaningful data for analysis depends on its development and core measurement properties, such as content validity, structural validity and internal consistency [17, 18].

The purpose of this study was to describe the development and psychometric properties of the items developed for a breast cancer recurrence instrument (BreastCaRe). We wished to maximize content validity and patient acceptability through the involvement of relevant stakeholders, i.e. clinical experts and patients. We further wished to assess structural validity (i.e. dimensionality based on items assessing symptoms of recurrence indicative of different organ sites), internal consistency, local independence and measurement invariance (absence of differential item functioning, DIF) to create a final measure that is suitable for further testing. A pair of items show local dependence (LD) if they are correlated beyond what the underlying latent variable can account for. This is important to evaluate as it may cause biased estimates of reliability. An item shows DIF when a variable systematically impacts responses in subgroups (e.g. when older women systematically score lower on the item than younger women who experience the same level of symptoms). This property is important to evaluate as it may cause biased results when comparing subgroups [17, 18]. Finally, we aimed to develop a scoring table to facilitate interpretation and potential implementation in the breast cancer follow-up clinic. Since the raw score metric is not in general equal to the latent variable metric, we used scoring

from Rasch analysis, which estimates item and person measures on the same scale.

Materials and methods

Context

This study is nested in the MyHealth randomized trial testing a nurse-led breast cancer follow-up program utilizing the systematic online collection of PRO in women who had completed curatively-intended primary breast cancer treatment [19]. The trial is ongoing and the items are used individually as part of a web-based screening tool, whereby a predefined algorithm alerts the nurse to contact a patient who has high scores on symptoms of late-effects or potential recurrence requiring attention [20]. The MyHealth trial was approved by the Ethics Committee of the Capital Region of Denmark (H-L6035885) and is registered on ClinicalTrials.gov (NCT02949167).

Development of items

We used current medical and clinical knowledge regarding the symptoms of recurrent breast cancer and how they vary depending on organ site [21]. We also reviewed the literature for available instruments used for the assessment of symptoms in cancer patients and noted the linguistic structure of the instructions and items. Based on this, we constructed different items that could potentially be relevant for the detection of symptoms of recurrence after breast cancer. These items were subsequently sent to six breast cancer specialists from the major oncology and breast cancer surgery departments in Denmark, who all accepted to participate in several rounds of semi-structured telephone-interviews with a researcher (LS).The most important symptoms and their relation to breast cancer recurrence sites were then selected and verified, and items were revised for clinical accuracy.

Scoring and hypothesized dimensions of the BreastCaRe

Based on a consensus from these interviews, 23 items were selected that assessed symptoms associated with recurrences in different organ sites. Instructions ask patients about any changes in physical symptoms within the past month. The first 22 items are answered on a four-point scale (0: Not at all; 1: A little; 2: Quite a bit; 3: Very much), while the final item was open-ended and allowed the patient to write down any other symptom causing her to worry about recurrent disease. Seven subscales were proposed based on clinical knowledge of symptoms that might indicate recurrence in general or in different organ sites: general symptoms (G: 5 items), bone recurrence (BO: 5 items), liver recurrence (LI: 4 items), lung recurrence (LU: 2 items), brain recurrence (BR: 5 items), locoregional recurrence (LO: 4 items) and contralateral recurrence (CO: 1 item). Three items (items 7, 13 and 14) were initially placed in more than one subscale, as the same symptom could be a sign of more than one type of recurrence (Table A).

Cognitive testing and patient involvement

In the MyHealth trial, a patient panel consisting of seven women who had completed primary treatment for breast cancer was established as part of efforts to involve patient representatives in the development of the trial. Procedures for the selection and involvement of patients in the development of the MyHealth trial have been reported elsewhere [22]. The panel was specially recruited such that half of the women did not have more than compulsory school education. The preliminary BreastCaRe items were tested by this panel of women and individual cognitive interviews were carried out based on an interview guide to assess acceptability and comprehensibility, item responses and instructions.

Test population

The BreastCaRe items were tested among breast cancer survivors. Between January and August 2017, 1773 breast cancer survivors affected by changes in routine follow-up after treatment for breast cancer were invited to participate in a questionnaire study. Invited patients had completed primary treatment between 2007 and 2015 at the Department of Clinical Oncology and Palliative Care, Zealand University Hospital, Denmark. They had attended routine specialist-led follow-up for 1–10 years, but were now being switched to either nurse-led or patient-initiated follow-up. Upon informed consent, a questionnaire was sent out at baseline and again after one year, including sociodemographic information and validated scales measuring other outcomes such as the European Organization for Research and Treatment of Cancer quality of life questionnaire (EORTC QLQ-C30) [23] in addition to the BreastCaRe items. This study is based on data from the baseline questionnaire. We further obtained information on vital status from The Danish Civil Registration System (CPR) [24] and on tumor characteristics and treatment from The Danish Breast Cancer Group (DBCG) [25].

Statistical analyses

Descriptive statistics were computed to examine the sociodemographic and clinical characteristics of both responders (defined as women who returned the questionnaire with informed consent) and non-responders in the study sample,

and the two groups were compared using chi-square tests. To obtain more robust and theory-driven results, we split the sample randomly into two subsets that were both sufficiently large for analysis. Measurement models were derived using CFA in the first randomly chosen subset and were then evaluated using Rasch analysis in the second subset. This permits investigation of the characteristics of items and persons, and allows the creation of a conversion table that translates ordinal observed scores to linear, interval-level estimates. Combining item pairs with LD into "single combination items" was used as a straightforward way of including LD in the model. This approach yielded the highest test information in a comparison with other methods [26 p.188].

Differential item functioning was assessed according to age group ($\leq 60, > 60$), education (low, medium, high), time since diagnosis (≤ 5 years, > 5 years), type of surgery (lumpectomy, mastectomy), lymph node dissection (yes, no), chemotherapy (yes, no), endocrine therapy (yes, no) and current health status (high/low). The Global Health Status score of the EORTC-C30 was used as a measure of health status [27]. The score was dichotomized into high versus low using a threshold of 66.7, which is the normative median score from the general population [27].

CFA models were fitted using complete cases, and both a six-dimensional CFA and a bifactor model were fitted. CFA for ordinal items based on polychoric correlations was done using the R package lavaan [28, 29]. The factors were derived theoretically, but to assess over-fitting the data, we also computed the empirical Kaiser criterion [30]. We report standardized factor loadings and associated standard errors together with the following fit statistics: Chi-square, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI). Items hypothesized to be on the same subscale should have similar and high loadings on that factor. Threshold levels for a good fit for a scale are indicated by chi-square p-value > 0.05, RMSEA ≤ 0.05 and CFI/TLI ≥ 0.95 [31]. Modification indices are used to improve the model by indicating items with local dependence. We report (i) CFA models without LD and (ii) CFA models where LD is included using latent variables, thus retaining the independence of error terms. In the CFA models DIF was evaluated using chi-square difference tests for nested models.

We used graphical Rasch models that can incorporate local dependence and DIF [32, 33] (details on the parametrization inference provided in supplementary materials). The similarity between the observed and expected responses to any item is reported through two fit statistics: (1) the outlier- sensitive fit statistic (OUTFIT) and (2) the information-weighted fit statistic (INFIT). Traditionally a fit value of 1.0 indicates perfect fit to model expectations and item fit values in the range between 0.7 and 1.3 are an acceptable indication of item fit [34, 35]. We used a conditional version of the fit statistics with known asymptotic properties [36] which makes it possible to compute P-values. Local dependence and DIF were evaluated using conditional likelihood ratio tests [37] and item screening [38]. We also evaluated the overall model fit for each subscales using Andersens conditional likelihood ratio test [39]. Furthermore, we did additional tests of DIF in the graphical Rasch model because this model is used to produce the scoring. A histogram of the item and person measure distributions plotted against each other (a Wright map) was created to evaluate if the items were well-targeted to the intended patient population. This was done for the total BreastCare scale and we also generate Wright maps for each of the six subscales based on the items in the subscale only.

The conversion table developed to convert the raw scores of each subscale to the metric of the latent variable was operationalized using Warms weighted likelihood estimates [40]. For ease of interpretation we also present a linear rescaling of these person measures to a zero to 100 scale. We do this for the total BreastCare scale and also generate conversion tables for each of the six subscales based on the items in the subscale only. The single-item contralateral recurrence subscale and the final item with the written response format were not included in the psychometric analyses. Analyses were carried out using SAS [41, 42] and DIGRAM [43]. All P-values were adjusted for multiple testing using the Benjamini–Hochberg correction.

Results

Content validity and cognitive testing

An initial list of twenty-two items was developed and sent to five clinical oncologists and one breast cancer surgeon from the main oncology/surgery departments in Denmark. Through telephone interviews and discussion, all the items were revised for clinical accuracy, while several items were deleted and replaced with new ones that were judged to be more relevant and comprehensive (Table 1). Individual cognitive interviews with the patient panel resulted in only minor adjustments in the wording of the instructions and individual items were made to make sure that they could be more accurately understood by patients (Table 1).

Test population

A total of 1178 (66.4%) questionnaires were returned with written informed consent and were included for analysis. Only small differences were seen between the two groups representing women included for analysis and those not included on clinical characteristics except for radiotherapy, where the study sample included a significantly higher

Table 1 Examples of revisions based on expert and patient feedback

Initial version example	Feedback	Revision example
	Expert feedback	
Tiredness Need to rest Dizziness	"*Unspecific,*" "*too general,*" "*Only important if it's new onset or persistent.*"	Items revised to include the phrases "more than usual" or "persistent"
Pain in breast area	"*Most likely a late-effect,*""*not usually indicative of recurrence.*"	Item deleted
Difficulty with controlling bowels/urinating	"*A late symptom of spinal cord compression (due to recurrence), ought to be detected earlier by symptoms of radiating pain and sensations in legs*."	Item deleted
Pain in upper abdomen	"*Too general, will catch patients with dyspepsia*."	Item replaced with "new pain and pressure in upper abdomen"
	Patient feedback	
Instructions requested patients to focus on health changes "since the last visit."	Patients found it difficult to remember when the last visit was and were not sure what to focus on	Instructions rephrased to "within the past month."
Radiating pain Changes in sensations in legs	Patients had difficulty understanding what type of pain or sensation were relevant	The examples "radiating from your back to your leg" and "tingling or numbness" were added to this items to aid patient understanding

percentage of women who received radiotherapy (Table 2). Responses to the BreastCaRe items were all right-skewed and the extent of missingness was low, ranging from 19 missing responses for item 15 (1.6%) to 33 missing responses for item 22 (2.8%) (Table 3).

Confirmatory factor analysis

We used information from the factor loadings to identify the best placement for the items that were initially placed on more than one subscale (Q7, Q13 and Q14). Item Q7 ('Lost weight without effort') was removed from the general symptoms of recurrence subscale and retained on the liver recurrence subscale. Item Q13 ('Reduced ability to lift your arm or leg') and Q14 ('Difficulty controlling the movement of arm or leg') were retained on the bone recurrence subscale and removed from the brain recurrence subscale.

For the six-dimensional CFA model (Fig. 1, panel (a)) the fit was adequate: chi-square = 278.7, df = 174, $p < 0.001$, RMSEA 0.031, (90% CI 0.024 to 0.038), CFI 0.988 and TLI 0.985. For the bi-factor model (Fig. 1, panel (b)) the fit was similar: chi-square = 306.3, df = 183, $p < 0.001$, RMSEA 0.033, (90% CI 0.026–0.039), CFI 0.986 and TLI 0.984. The six-dimensional CFA confirmed the hypothesized structure, but since the empirical Kaiser criterion suggested three factors the more parsimonious bi-factor model may be preferable.

Modification indices indicated the presence of five locally dependent item pairs: Q1 ('Felt more tired than usual') and Q2 ('Had to rest more than usual'); Q3 ('Felt sick or unwell') and Q6 ('Nausea'); Q4 ('Newly developed bone pain') and Q11 ('Experienced pain that radiates'); Q5 ('Appetite decreased') and Q7 ('Lost weight without effort'); and Q13 ('Reduced ability to lift arm or leg') and Q14 ('Difficulty controlling movement of arm or leg'). A clinical decision was made to retain all the items. Adding additional latent variables to the model to account for the additional correlation yielded a six-dimensional CFA model with better fit: chi-square = 185.7, df = 169, $p = 0.180$, RMSEA 0.013, (90% CI 0.000 to 0.023), CFI 0.998 and TLI 0.998. The fit of the bi-factor CFA model also improved (chi-square = 200.0, df = 178, $p = 0.124$, RMSEA 0.014, (90% CI 0.000 to 0.023), CFI 0.997 and TLI 0.997). The factor structure of the models with additional latent variables to the model to account for the additional correlation is reported in Fig. 1. No evidence of DIF was identified by the multiple groups CFA (Table 4).

Rasch analysis

For all six subscales, the models derived using CFA were confirmed using graphical Rasch models and indicated acceptable item fit. Table 3 contains the estimated item threshold parameters, and the conditional infit and outfit item fit statistics. The item fit statistics evaluate item fit in the six subscales. For the general symptoms of recurrence subscale, the bone recurrence subscale, and the liver recurrence subscale, fit to graphical Rasch models taking into account item local dependence was confirmed. For the lung, brain and locoregional recurrence subscales, fit to Rasch models without item local dependence was confirmed. Overall Rasch model fit was acceptable for all subscales (Table 5).

Table 2 Characteristics of the test population of breast cancer survivors in this study comparing respondents and non-respondents

	Respondents $n=1178$	Non-respondents $n=595$	p-value
Age, N (%)	267 (22.7)	154 (25.9)	0.13
≤60 years	911 (77.3)	441 (74.1)	
>60 years			
Education, N (%)	456 (38.7)	N/A	
Short	227 (19.3)		
Medium	422 (36.0)		
Long	73 (6.2)		
Unknown			
Time since diagnosis, N (%)	510 (43.3)	307 (51.6)	0.20
≤5 years	654 (55.5)	283 (47.6)	
>5 years	14 (1.2)	5 (0.8)	
Unknown			
Type of surgery, N (%)	900 (76.4)	425 (71.4)	0.07
Lumpectomy	242 (20.5)	145 (24.4)	
Mastectomy	36 (3.1)	25 (4.2)	
Unknown			
Axillary dissection, N (%)	487 (41.3)	233 (39.1)	0.68
Yes	689 (58.5)	361 (60.7)	
No	2 (0.2)	1 (0.2)	
Unknown			
Trastuzumab, N (%)	100 (8.5)	48 (8.1)	0.88
Yes	986 (83.7)	497 (83.5)	
No	92 (7.8)	50 (8.4)	
Unknown			
Chemotherapy, N (%)	443 (37.6)	208 (35.0)	0.45
Yes	580 (49.2)	299 (50.3)	
No	155 (13.2)	88 (14.8)	
Unknown			
Radiotherapy, N (%)	974 (82.7)	451 (75.8)	0.003*
Yes	15 (1.3)	11 (1.8)	
No	189 (16.0)	133 (22.4)	
Unknown			
Endocrine therapy, N (%)	885 (75.1)	426 (71.6)	0.23
Yes	158 (13.4)	86 (14.5)	
No	135 (11.5)	83 (13.9)	
Unknown			
Global health status, N (%)	445 (37.8)	N/A	
Low	720 (61.1)		
High	13 (1.1)		
Unknown			

Histograms of item and person measures (Wright maps) illustrate that the measure has a floor effect, but that items are located throughout the range of person measure estimate values, indicating that item set is well-targeted to the high end of the intended patient population (Fig. 2). This means person measures can be estimated with reasonable precision towards the high end of the scale. The same is true for Wright maps for the six subscales (included in the supplementary material), but of course the precision is smaller in shorter subscale. The conversion tables illustrate that differences between consecutive raw scores do not represent equal intervals on the latent variable axis (Table 6; conversion tables for the six subscales included in the supplementary material).

Although CFA indicated no evidence of DIF, analyses using graphical Rasch models indicated some evidence of DIF for the items Q3 ('Felt sick or unwell') and Q15 ('Strong, persistent headaches') with respect to global health status, and item Q22 ('Swelling in arm') with respect to axillary dissection. None of these findings were significant after adjustment for multiple testing and the magnitude of DIF was quite small (results available upon request). For this reason, they were not incorporated into the current conversion tables.

Table 3 Distribution of answers and psychometric properties of BreastCaRe items and subscales ($n = 1178$)

sub-sca-le[a]	Item		N	Not at all %	A little %	Quite a bit	Very much	Factor loading (S.E.)[b]	Rasch analysis						
									Item thresholds			In-fit[c]	p-value	Out-fit[c]	*p*-value
G	1	Felt more tired than usual	1152	63	25	9	3	0.864 (0.029)	− 1.64	− 0.56	0.01	1.03	0.7031	1.05	0.7762
	2	Had to rest more than usual	1154	66	24	8	3	0.883 (0.026)	− 1.50	− 0.38	0.06	0.97	0.7922	0.96	0.8298
	3	Felt sick or unwell	1147	72	21	6	1	0.842 (0.028)	− 1.13	− 0.16	0.71	1.02	0.6942	1.04	0.8507
BO	4	Newly developed bone pain	1146	57	25	12	6	0.761 (0.033)	− 1.87	− 1.08	-0.51	1.02	0.1948	1.10	0.7701
	11	Pain that radiates	1147	72	18	6	3	0.680 (0.041)	− 1.03	− 0.44	-0.11	0.99	0.9309	1.01	0.9279
	12	Changes in sensations in legs	1151	77	14	6	3	0.767 (0.035)	− 0.65	− 0.63	0.01	0.96	0.3382	0.90	0.6868
	13	Reduced ability to lift arm or leg	1155	67	21	10	2	0.865 (0.030)	− 1.33	− 0.78	0.46	1.01	0.7437	1.02	0.8552
	14	Difficulty controlling movement of arm or leg	1156	89	7	2	1	0.691 (0.057)	0.37	− 0.01	0.27	1.04	0.7288	0.89	0.7353
LI	5	Appetite decreased	1157	83	13	3	1	0.713 (0.046)	− 0.31	0.11	0.47	0.96	0.6725	0.96	0.7152
	6	Nausea	1156	88	10	2	0	0.591 (0.052)	0.07	0.49	1.12	0.85	0.2391	0.83	0.3023
	7	Lost weight without effort	1156	88	8	2	1	0.628 (0.059)	0.20	0.16	− 0.19	1.07	0.5786	1.07	0.6210
	8	New pain or pressure in upper abdomen	1154	84	12	3	1	0.644 (0.053)	− 0.21	− 0.03	0.74	1.16	0.1920	1.10	0.1281
LU	9	Become short of breath faster than usual	1155	64	28	6	2	0.855 (0.046)	− 1.65	− 0.05	0.23	0.99	0.9557	1.00	0.9760
	10	Persistent cough	1155	84	12	4	1	0.625 (0.050)	− 0.23	− 0.11	0.54	0.99	0.9557	1.00	0.9760
BR	15	Strong, persistent headaches	1159	91	6	2	1	0.769 (0.058)	0.67	− 0.16	0.21	1.07	0.8457	1.03	0.7858
	16	Persistent dizziness	1155	83	12	4	1	0.871 (0.040)	− 0.29	− 0.08	0.32	0.91	0.2794	0.92	0.3417
	17	Visual disturbances	1154	79	16	4	1	0.656 (0.050)	− 0.67	0.03	0.66	1.02	0.9732	1.00	0.8306
LO	18	Lumps in operated breast	1148	95	3	1	0	0.608 (0.100)	1.33	0.30	1.98	0.94	0.6629	0.94	0.7254
	20	Changes in skin around scar	1154	87	10	3	0	0.566 (0.078)	− 0.01	0.20	1.18	1.09	0.3437	1.04	0.2817
	21	Lumps on throat or in armpit	1154	97	2	1	0	0.775 (0.127)	1.76	0.25	1.70	0.91	0.9465	0.98	0.7274
	22	Swelling in arm	1145	91	7	2	1	0.659 (0.086)	0.48	0.09	0.61	1.12	0.6377	1.04	0.2247
CO	19	Lumps in opposite breast	1150	98	2	0	0	–	–	–	–	–	–	–	–

[a]*G* general symptoms, *BO* bone recurrence, *LI* liver recurrence, *LU* lung recurrence, *BR* brain recurrence, *LO* loco-regional recurrence

[b]S.E.: standard error

[c]The similarity between the observed and expected responses to any item is reported through the outlier- sensitive fit statistic (OUTFIT) and the information-weighted fit statistic (INFIT). Fit value of 1.0 indicates perfect fit to model expectations and item fit values in the range between 0.7 and 1.3 are an acceptable indication of item fit. *p*-values < 0.05 indicate significant difference between the observed and expected values

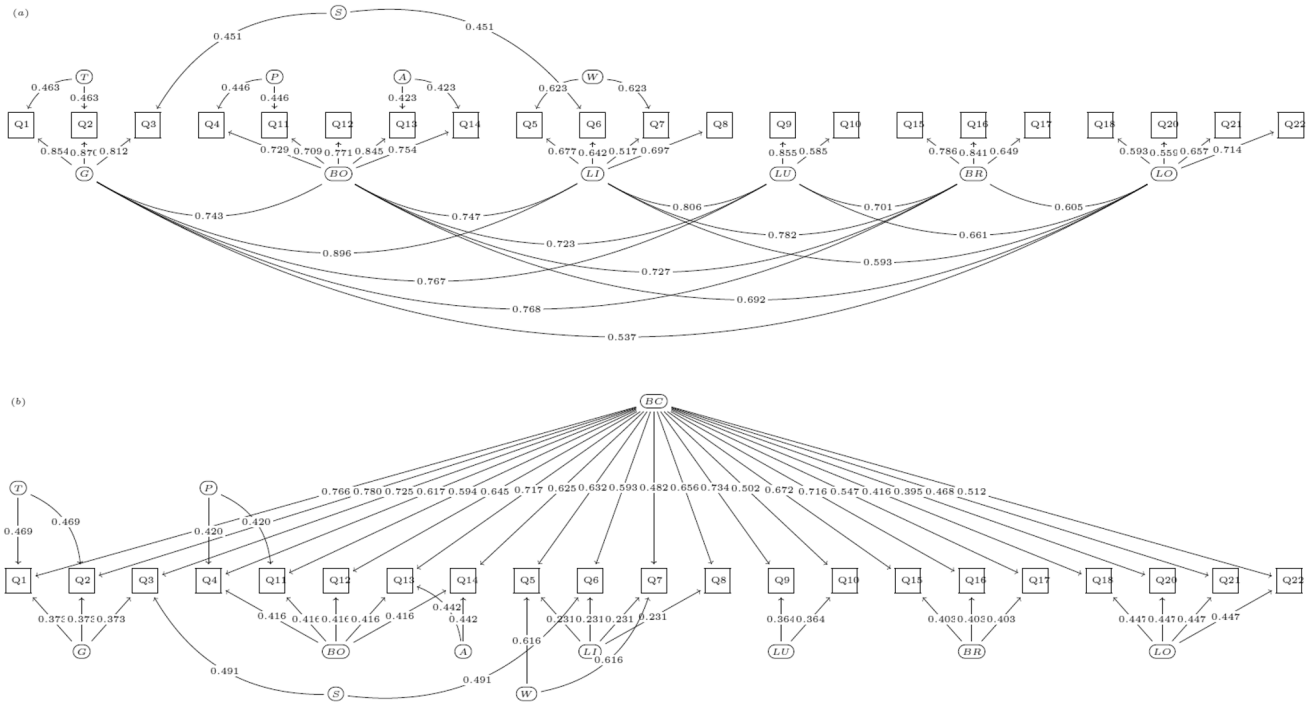

Fig. 1 The factor structure of the BreastCaRe instrument taking local dependence into account. Panel (a) shows the six-dimensional CFA model, panel (b) show the bifactor model. *BC* BreastCare total symptom, *G* general symptoms of recurrence, *BO* bone recurrence, *LI* liver recurrence, *LU* lung recurrence, *BR* brain recurrence, *LO* loco-regional recurrence, *T* Tired/rest, *P* Pain, *A* Arm/leg, *S* Sick/nausea, *W* Weight. Observed variables are shown in boxes, factors and bi-factors are shown in circles

Table 4 Tests of Measurement invariance (absence of DIF)

Variable	Chi-square difference test for nested models		
	χ^2	df	*p*
Age (≤ 60 years, > 60 years)	16.6	15	0.34
Education (Short, Medium, Long)	37.8	30	0.16
Time since diagnosis (≤ 5 years, > 5 years)	17.3	15	0.30
Type of surgery (Lumpectom; Mastectomy)	18.1	15	0.26
Axillary dissection (yes; no)	19.6	15	0.19
Chemotherapy (yes; no)	14.6	15	0.48
Endocrine therapy (yes; no)	17.0	15	0.32
Global health status, (high; low)	18.9	15	0.22

Table 5 Evaluation of over-all Rasch model fit for the BreastCare subscales

Domain	Andersens conditional likelihood ratio test		
	χ^2	df	*p*
G: general symptoms	8.6	14	0.86
BO: bone recurrence	16.6	24	0.87
LI: liver recurrence	2.4	16	1.00
LU: lung recurrence	8.2	5	0.14
BR: brain recurrence	1.2	8	1.00
LO: loco-regional recurrence	14.7	9	0.10

Discussion

We aimed for the first time to develop and investigate the psychometric properties of a new PRO instrument to screen for symptoms of recurrence after breast cancer (BreastCaRe). We involved clinical experts and patients to maximize the instrument's content validity, patient acceptability and comprehension, and tested structural validity, internal consistency and measurement invariance. We ended with 22 multiple-choice items falling on seven distinct subscales, with one subscale consisting of a single item. One free-text item was included to create a final 23-item instrument and a scoring table based on Rasch analysis was developed for further measurement evaluation. We are now preparing to examine the predictive validity of this instrument using recurrence data from the study population as the next step in validating the BreastCaRe.

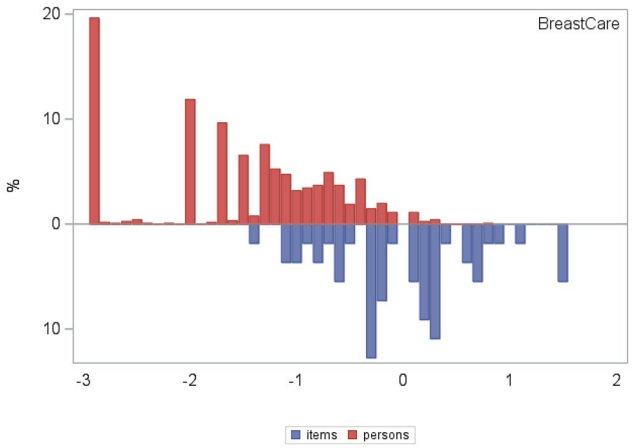

Fig. 2 Wright maps for the BreastCare instrument. Person measures above the vertical axis and item measures below the vertical axis

In the currently proposed factor structure, three items originally repeated on more than one subscale based on clinical knowledge were placed on the subscales based on psychometric fit. A bi-factor model also fitted the data adequately. Both models also required local dependence for five item pairs to get optimal fit. This indicates potential redundancy and a possible solution is to remove items to create a more parsimonious measure [17]. However, we kept all the items as the clinical experts judged them to provide different and important clinical information. Rasch models taking local dependence into account were used to develop a conversion scoring table where the raw ordinal scores of each subscale are converted to standardized interval scores ranging from zero to 100. We present these for the total BreastCare scale and for each of the six subscales to help clinicians interpret scores and changes in scores over time, but stress that (i) a score of zero (or 100) indicates that a patient has scored in the lowest (highest) category on all items in the scale or subscale, but that it is possible to be located lower (higher) and that (ii) these conversion tables are only relevant in populations where the item calibrations presented here are deemed relevant.

The Wright maps illustrate that items are well-targeted to the intended patient population and that person measures can also be estimated towards the high end of the scales. However for the short subscales the measurement precision may not be sufficient. The conversion tables illustrate that differences between consecutive raw scores do not correspond to equal intervals on the latent variable axis, which is why raw scores should not be used in parametric analyses. This is rectified using Rasch analysis which estimates item and person measures on the same scale. Simple conversion tables result from using Rasch analysis, where the total score is sufficient and the ability of the graphical Rasch model of incorporating local dependence makes it a natural choice in this setting.

We also investigated DIF with regards to age, overall health status, surgery and adjuvant treatments received, which is important if the instrument is used for comparisons because DIF can bias comparisons of groups or individuals [44]. DIF was assessed by evaluating the fit of multiple groups CFA models and this is only feasible when the groups are large enough. Thus the potential effects of two adjuvant treatment modalities (Trastuzumab and radiotherapy) on DIF

Table 6 Conversion table for scoring the BreastCaRe instrument

Sumscore[a]	θ	S	Sumscore	θ	S	Sumscore	θ	S	Sumscore	θ	S
0	− 1.884	0.0	16	1.765	40.7	32	2.674	50.9	48	3.568	60.9
1	− 0.774	12.4	17	1.834	41.5	33	2.724	51.5	49	3.639	61.7
2	− 0.253	18.2	18	1.901	42.3	34	2.774	52.0	50	3.715	62.5
3	0.093	22.1	19	1.966	43.0	35	2.825	52.6	51	3.795	63.4
4	0.353	25.0	20	2.028	43.7	36	2.876	53.2	52	3.880	64.4
5	0.562	27.3	21	2.088	44.4	37	2.927	53.7	53	3.973	65.4
6	0.738	29.3	22	2.146	45.0	38	2.979	54.3	54	4.074	66.5
7	0.889	31.0	23	2.203	45.6	39	3.031	54.9	55	4.185	67.8
8	1.023	32.5	24	2.259	46.3	40	3.085	55.5	56	4.310	69.2
9	1.144	33.8	25	2.313	46.9	41	3.139	56.1	57	4.452	70.8
10	1.253	35.0	26	2.367	47.5	42	3.195	56.7	58	4.619	72.6
11	1.353	36.2	27	2.419	48.1	43	3.252	57.4	59	4.820	74.9
12	1.447	37.2	28	2.471	48.6	44	3.311	58.0	60	5.075	77.7
13	1.533	38.2	29	2.522	49.2	45	3.372	58.7	61	5.419	81.6
14	1.615	39.1	30	2.573	49.8	46	3.434	59.4	62	5.946	87.4
15	1.692	39.9	31	2.624	50.3	47	3.500	60.1	63	7.070	100.0

[a]Sum of the raw scores of all the items on the instrument

Estimated person measure (θ), Standardized scores (S)

were not studied, as very few patients received Trastuzumab and almost all received radiotherapy. We found no evidence of DIF in the CFA analysis, but Rasch analysis indicated that: (i) the items Q3 ('Have you felt sick or unwell?') and Q17 ('strong persistent headaches') could potentially be biased in comparisons of respondents with different levels of global health status, and that (ii) the item Q22 ('Swelling in arm') could potentially be biased in comparisons between respondents who received axillary dissection and respondents who did not. However, these findings were not significant after adjustment for multiple testing. This potential presence of DIF should be investigated further and perhaps be incorporated into conversion tables.

Strengths and limitations

The involvement of clinical experts and patients maximized content validity, acceptability and comprehensibility of this instrument, thus enhancing future uptake and applicability. A further strength of this study is the large size of the test population, which allowed us to use two methods to investigate the psychometric properties by splitting our sample randomly in two. The Rasch analysis confirmed the results of the factor analysis, producing a final instrument with good psychometric properties. We also identified items that were associated with each other (local dependence) and incorporated this information in a conversion table for use in further validation analyses and potential clinical application.

As items were developed based on clinical consensus, we cannot rule out the possibility that other symptoms relevant for the detection of breast cancer recurrence may have been overlooked. However, BreastCaRe items capture the most common symptoms presented at the time of recurrence detection, as reported in a recent cross-sectional study of 310 breast cancer patients [14]. The response rate of 66.4% means that we cannot exclude problems with the generalizability of the study population, although differences between the respondents and non-respondents were generally small. Also, the low levels of symptoms in this population and lack of variability in the data can be a challenge for statistical models, but the CFA for ordinal items based on polychoric correlations converged without problems. Finally, a drawback of using polytomous Rasch models is that they are logically inconsistent with the assumption that a rating scale with ordered thresholds is used to rate items. The method of successive dichotomizations (MSD) [45] or the graded response model [46] might be a better choice for a scoring algorithm, but taking local dependence into account is not straight-forward in these models. Where results were found to differ between the CFA and Rasch analysis this is likely due to the differences in the way these two approaches models means (locations) vs variances (discriminations). However, a major strength of the Rasch models is that item fit statistics with known asymptotic distributions can be used [36, 47].

Early detection and improved treatment of breast cancer have led to a tremendous increase in the number of breast cancer survivors and the subsequent challenge of allocating resources for follow-up in this group of patients [48, 49]. The fact that breast cancer may relapse many years after primary treatment points to the need for more cost-effective and convenient surveillance methods such as the use of PROMs. To our knowledge, no trial has yet investigated the effect of a follow-up program based on PRO monitoring of symptoms of recurrence on the outcomes of survival or detection of recurrence after primary breast cancer treatment and we lack a PRO measure for detection of recurrence. Items that make up the BreastCaRe are currently being used in the MyHealth randomized trial investigating a nurse-led breast cancer follow-up program based on systematic PRO monitoring, [20] and results from this trial will inform the future investigation of this instrument.

Conclusion

The BreastCaRe is a psychometrically sound, recurrence-specific PROM developed for monitoring women who have completed primary breast cancer treatment. If the predictive validity of the BreastCaRe is confirmed in our next step analyses, it has the potential to be a valuable and cost-effective tool in the organization of future breast cancer follow-up.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11136-021-02841-1.

Acknowledgements We wish to thank our expert panel (senior consultants Dr Birgitte Offersen, Dr Michael Andersson, Dr Niels Kroman, and Dr Anders Bonde), our patient panel, and Julie A. Skaarup, who carried out the patient cognitive interviews. We also thank Visti B. Larsen for his invaluable assistance in data management.

Author contributions All authors contributed to the study conception and design, data collection, analysis and/or interpretation. All authors contributed to and approved the final manuscript.

Funding The MyHealth trial was funded by the Danish Cancer Society, Region Zealand, Rigshospitalet (Copenhagen University Hospital) and the Capital Region of Denmark.

Data Availability The data that support the findings of this study are not publicly available due to data protection legislation.

Code availability Available upon request.

Compliance with ethical standards

Conflict of interest All authors declare no conflict of interest.

References

1. Yang, L. Y., Manhas, D. S., Howard, A. F., & Olson, R. A. (2018). Patient-reported outcome use in oncology: a systematic review of the impact on patient-clinician communication. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer., 26*(1), 41–60. https://doi.org/10.1007/s00520-017-3865-7.
2. Basch, E. M., Deal, A. M., Dueck, A. C., Bennett, A. V., Atkinson, T. M., Scher, H. I., et al. (2017). Overall survival results of a randomized trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. *Journal of Clinical Oncology, 318*(2), 197–198. https://doi.org/10.1200/JCO.2017.35.18_suppl.LBA2.
3. Chen, J., Ou, L., & Hollis, S. J. (2013). A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. *BMC Health Services Research, 13*, 211. https://doi.org/10.1186/1472-6963-13-211.
4. Runowicz, C. D., Leach, C. R., Henry, N. L., Henry, K. S., Mackey, H. T., Cowens-Alvarado, R. L., et al. (2016). American Cancer Society/American Society of Clinical Oncology breast cancer survivorship care guideline. *CA: A Cancer Journal for Clinicians, 66*(1), 43–73. https://doi.org/10.3322/caac.21319.
5. Howell, D., Molloy, S., Wilkinson, K., Green, E., Orchard, K., Wang, K., et al. (2015). Patient-reported outcomes in routine cancer clinical practice: a scoping review of use, impact on health outcomes, and implementation factors. *Annals of Oncology, 26*(9), 1846–1858. https://doi.org/10.1093/annonc/mdv181.
6. Coulter A, Potter C, Peters M, Fitzpatrick R. Cancer PROMs: a scoping study: Macmillan Cancer Support2015.
7. WHO. Breast cancer burden. 2018. http://www.who.int/cancer/detection/breastcancer/en/index1.html. Accessed 08/08/2018.
8. Colleoni, M., Sun, Z., Price, K. N., Karlsson, P., Forbes, J. F., Thürlimann, B., et al. (2016). Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: results from the international breast cancer study group trials I to V. *Journal of Clinical Oncology, 34*(9), 927–935. https://doi.org/10.1200/JCO.2015.62.3504.
9. Pan, H., Gray, R., Braybrooke, J., Davies, C., Taylor, C., McGale, P., et al. (2017). 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *New England Journal of Medicine, 377*(19), 1836–1846. https://doi.org/10.1056/NEJMoa1701830.
10. Ahmad, A. (2013). Pathways to breast cancer recurrence. *ISRN Oncol, 2013*, 16. https://doi.org/10.1155/2013/290568.
11. Wu, X., Baig, A., Kasymjanova, G., Kafi, K., Holcroft, C., Mekouar, H., et al. (2016). Pattern of local recurrence and distant metastasis in breast cancer by molecular subtype. *Cureus., 8*(12), e924. https://doi.org/10.7759/cureus.924.
12. Taggart, F., Donnelly, P., & Dunn, J. (2012). Options for early breast cancer follow-up in primary and secondary care - a systematic review. *BMC Cancer, 12*(1), 238. https://doi.org/10.1186/1471-2407-12-238.
13. Pivot, X., Asmar, L., Hortobagyi, G. N., Theriault, R., Pastorini, F., & Buzdar, A. (2000). A retrospective study of first indicators of breast cancer recurrence. *Oncology, 58*(3), 185–190. https://doi.org/10.1159/000012098.
14. Saltbæk, L., Horsboel, T. A., Offersen, B. V., Andersson, M., Friberg, A. S., Skriver, S. K., et al. (2020). Patterns in detection of recurrence among patients treated for breast cancer. *Breast Cancer Research and Treatment*. https://doi.org/10.1007/s10549-020-05847-4.
15. Denis, F., Lethrosne, C., Pourel, N., Molinier, O., Pointreau, Y., Domont, J., Bourgeois, H., Senellart, H., Trémolières, P., Lizée, T., Bennouna, J., Urban, T., El Khouri, C., Charron, A., Septans, A.-L., Balavoine, M., Landry, S., Solal-Céligny, P., & Letellier, C. (2017). Randomized trial comparing a web-mediated follow up with routine surveillance in lung cancer patients. *Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/djx029.
16. Denis, F., Viger, L., Charron, A., Voog, E., Dupuis, O., Pointreau, Y., et al. (2014). Detection of lung cancer relapse using self-reported symptoms transmitted via an Internet Web-application: pilot study of the sentinel follow-up. *Supportive Care in Cancer, 22*(6), 1467–1473. https://doi.org/10.1007/s00520-013-2111-1.
17. Dima, A. L. (2018). Scale validation in applied health research: tutorial for a 6-step R-based psychometrics protocol. *Health Psychology and Behavioral Medicine., 6*(1), 136–161. https://doi.org/10.1080/21642850.2018.1472602.
18. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation., 19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8.
19. ClinicalTrials.gov. MyHealth: Follow-up After Breast Cancer Treatment (MyHealth). 2016. https://clinicaltrials.gov/ct2/show/NCT02949167.
20. Saltbæk, L., Karlsen, R. V., Bidstrup, P. E., Høeg, B. L., Zoffmann, V., Horsbøl, T. A., et al. (2019). MyHealth: specialist nurse-led follow-up in breast cancer. A randomized controlled trial – development and feasibility. *Acta Oncologica*. https://doi.org/10.1080/0284186X.2018.1563717.
21. Bast, R. C., Croce, C. M., Hait, W. N., Hong, W. K., Kufe, D. W., Piccart-Gebhart, M., et al. (Eds.). (2017). *Cancer Medicine*. (9th ed.). Wiley-Blackwell: Hoboken, NJ.
22. Høeg, B. L., Tjørnhøj-Thomsen, T., Skaarup, J. A., Langstrup, H., Zoffmann, V., Saltbaek, L., et al. (2019). Whose perspective is it anyway? Dilemmas of patient involvement in the development of a randomized clinical trial – a qualitative study. *Acta Oncologica, 58*(5), 1–8. https://doi.org/10.1080/0284186X.2019.1566776.
23. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute, 85*(5), 365–376.
24. Pedersen, C. B., Gøtzsche, H., Møller, J. Ø., & Mortensen, P. B. (2006). The Danish civil registration system. A cohort of eight million persons. *Danish Medical Bulletin, 53*, 441–449.
25. Jensen, M. B., Laenkholm, A. V., Offersen, B. V., Christiansen, P., Kroman, N., Mouridsen, H. T., et al. (2018). The clinical database and implementation of treatment guidelines by the Danish Breast Cancer Cooperative Group in 2007–2016. *Acta Oncologica, 57*(1), 13–18. https://doi.org/10.1080/0284186x.2017.1404638.
26. Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178–194. https://doi.org/10.1177/0146621616677520.
27. Scott NW, Fayers PM, Aaronson PM, Bottomley A, de Graeff A, Groenvold A et al. EORTC QLQ-C30 Reference Values 2008.
28. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software; Vol 1, Issue 2 (2012). 2012.
29. Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using mplus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal., 27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776.

30. Braeken, J., & van Assen, M. (2017). An empirical Kaiser criterion. *Psychological Methods, 22*(3), 450–466. https://doi.org/10.1037/met0000074.
31. Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *The Journal of Educational Research., 99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338.
32. Christensen KB, Kreiner S, Mesbah M, editors. Rasch Models in Health. Wiley; 2012.
33. Kreiner, S., & Christensen, K. B. (2002). Graphical Rasch models. In M. Mesbah, B. F. Cole, & M. L. T. Lee (Eds.), *Statistical Methods for Quality of Life Studies: Design, Measurements and Analysis.* (pp. 187–203). Boston: Springer.
34. Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*(1), 66–78.
35. Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*, 33. https://doi.org/10.1186/1471-2288-8-33.
36. Christensen K, Kreiner S. Item Fit Statistics. In: Christensen K, Kreiner S, Mesbah M, editors. Rasch Models in Health. 2013. p. 83–104.
37. Kreiner, S., & Christensen, K. B. (2013). Two Tests of Local Independence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models in Health Hoboken.* (pp. 131–136). NJ: John Wiley & Sons.
38. Kreiner, S., & Christensen, K. B. (2011). Item screening in graphical loglinear rasch models. *Psychometrika, 76*(2), 228–256. https://doi.org/10.1007/s11336-011-9203-y.
39. Kreiner S, Christensen KB. Overall Tests of the Rasch Model. In: Kreiner S, Christensen KB, Mesbah M, editors. Rasch Models in Health. 2013. p. 105–10.
40. Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627.
41. Christensen, K. B. (2013). Conditional maximum likelihood estimation in polytomous rasch models using SAS. *ISRN Computational Mathematics, 2013*, 1–8.
42. Christensen, K. B. (2006). Fitting polytomous Rasch models in SAS. *Journal of Applied Measurement, 7*(4), 407–417.
43. Kreiner S, Nielsen T. Item analysis in DIGRAM 3.04: Part I: Guided tours. Department of Biostastistics, University of Copenhagen 2013.
44. Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE Life Sciences Education*. https://doi.org/10.1187/cbe.16-10-0307.
45. Bradley, C., & Massof, R. W. (2018). Method of successive dichotomizations: An improved method for estimating measures of latent variables from rating scale data. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0206106.
46. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(1), 1–97. https://doi.org/10.1007/BF03372160.
47. Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications., 7*(1), 5. https://doi.org/10.1186/s40488-020-00108-7.
48. Montgomery, D. A., Krupa, K., & Cooke, T. G. (2007). Alternative methods of follow up in breast cancer: a systematic review of the literature. *British Journal of Cancer, 96*(11), 1625–1632. https://doi.org/10.1038/sj.bjc.6603771.
49. DeSantis, C., Ma, J., Bryan, L., & Jemal, A. (2014). Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians, 64*(1), 52.