Qual Life Res
DOI 10.1007/s11136-014-0658-z

Evaluation properties of the French version of the OUT-PATSAT35 satisfaction with care questionnaire according to classical and item response theory analyses

M. Panouillères · A. Anota · T. V. Nguyen · A. Brédart · J. F. Bosset · A. Monnier · M. Mercier · J. B. Hardouin

Accepted: 20 February 2014

Abstract
Purpose The present study investigates the properties of the French version of the OUT-PATSAT35 questionnaire, which evaluates the outpatients' satisfaction with care in oncology using classical analysis (CTT) and item response theory (IRT).
Methods This cross-sectional multicenter study includes 692 patients who completed the questionnaire at the end of their ambulatory treatment. CTT analyses tested the main psychometric properties (convergent and divergent validity, and internal consistency). IRT analyses were conducted separately for each OUT-PATSAT35 domain (the doctors, the nurses or the radiation therapists and the services/organization) by models from the Rasch family. We examined the fit of the data to the model expectations and tested whether the model assumptions of unidimensionality, monotonicity and local independence were respected.
Results A total of 605 (87.4 %) respondents were analyzed with a mean age of 64 years (range 29–88). Internal consistency for all scales separately and for the three main domains was good (Cronbach's α 0.74–0.98). IRT analyses were performed with the partial credit model. No disordered thresholds of polytomous items were found. Each domain showed high reliability but fitted poorly to the Rasch models. Three items in particular, the item about "promptness" in the doctors' domain and the items about "accessibility" and "environment" in the services/organization domain, presented the highest default of fit. A correct fit of the Rasch model can be obtained by dropping these items. Most of the local dependence concerned items about "information provided" in each domain. A major deviation of unidimensionality was found in the nurses' domain.
Conclusions CTT showed good psychometric properties of the OUT-PATSAT35. However, the Rasch analysis revealed some misfitting and redundant items. Taking the above problems into consideration, it could be interesting to refine the questionnaire in a future study.

Electronic supplementary material The online version of this article (doi:10.1007/s11136-014-0658-z) contains supplementary material, which is available to authorized users.

M. Panouillères · A. Anota · T. V. Nguyen · J. F. Bosset · M. Mercier
EA3181, University of Franche-Comte, Besançon, France

M. Panouillères (✉) · A. Anota
Methodological and Quality of Life in Oncology Unit, University Hospital of Besançon, Besançon, France
e-mail: marie.panouilleres@gmail.com

A. Anota · M. Mercier
Quality of Life in Oncology Clinical Research Platform, Besançon, France

A. Brédart
Psycho-Oncology Unit, Institut Curie, Paris, France

J. F. Bosset
Oncology-Radiotherapy Department, Besançon University Hospital, Besançon, France

A. Monnier
Radiotherapy Department, Montbeliard Hospital, Montbeliard, France

J. B. Hardouin
EA4275-SPHERE, University of Nantes, Nantes, France

J. B. Hardouin
Unit of Biostatistics and Methodology, University Hospital of Nantes, Nantes, France

Published online: 07 March 2014

Keywords Satisfaction with care · OUT-PATSAT35 questionnaire · Cancer · Item response theory · Classical test theory

Abbreviations

AIC	Akaike information criterion
CT	Chemotherapy
CTT	Classical test theory
DIF	Differential item functioning
EFA	Exploratory factor analysis
EORTC	European Organization for Research and Treatment of Cancer
HRQoL	Health-related quality of life
IRT	Item response theory
PCA	Principal component analysis
PCM	Partial credit model
PRO	Patient-reported outcomes
SC	Satisfaction with care
SD	Standard deviation
RSM	Rating scale model
RT	Radiotherapy

Background

Over the last two decades, a growing interest for patient-reported outcomes (PRO) evaluated through self-reported questionnaires has emerged as in the field of health-related quality of life (HRQoL). Developed in health outcome research and clinical practice, PRO generally covers concepts of patient satisfaction with care (SC), adherence to treatment and symptomatic to functional status [1]. PRO measurement in clinical settings guides treatment planning, management and monitoring. Thus, patient SC became increasingly important. Beyond the legal requirement to monitor the quality of hospital and the care areas, SC is an important indicator, especially in oncology, as it contributes to the assessment of the quality of care, influences a patient's adherence and impacts on outcomes [2, 3].

To date and despite the general excitement about the development of optimal HRQoL instruments and PRO assessment tools, only few reports of development and validation of French SC questionnaires exist in ambulatory oncology.

The Quality of Life working group of the European Organization for Research and Treatment of Cancer (EORTC) developed the IN-PATSAT32 questionnaire to assess inpatient SC in oncology units. In 2005, an international multicenter study validated it [4]. This questionnaire was translated and adapted in many countries and can be used regardless of the tumor location. It evaluates both the multidisciplinary care teams (hospital doctors and nurses) and health facility (care organization and services). However, the evolution of cancer treatment to outpatient treatment required the development of a new questionnaire with more specific aspects of ambulatory, not found in IN-PATSAT32 questionnaire: accessibility, continuity and coordination of care, and outpatient environment.

Based on the EORTC IN-PATSAT32, the OUT-PATSAT35 questionnaire has been developed in French [5] and is designed to assess the perception of outpatients on the quality of the care they have received. Developed in French language and also adapted in Spanish, its psychometric properties were determined according to classical test theory (CTT). Adequate psychometric properties have been found [5–7]. However, divergent validity was not respected by several items, especially in the domain evaluating the services and care organization. Thus, the authors mentioned the possibility to reconsider the number of dimensions of the instrument and to also achieve reducing the number of items. Furthermore, even if the results of these traditional CTT approaches were overall satisfactory, this analysis is insufficient as it did not consider some important aspects of measurement such as quality of measure's targeting, item properties estimated independently of the sample (difficulty, categories structure) and relevance of each item separately in the questionnaire's construct.

To enhance the quality of the assessment tools for health status outcomes, new and modern psychometric approaches using item response theory (IRT) modeling have emerged just over a decade ago and have notably progressed over the past 5 years. IRT models are statistical models establishing that the probability of a given respondent to an item depends on the item characteristics and on the respondent level on the construct being measured by the scale (SC or HRQoL as example, generally called the latent trait) [8]. To date, many authors confirm that IRT models are beneficial in overcoming the short outcomes of CTT, even if they do not replace the traditional psychometric analysis, and that these models are powerful and can be considered as supplementary tools.

Even though the OUT-PATSAT35 instrument has been psychometrically tested, using CTT, its robustness has not been investigated by IRT analysis. The goal of the present study was then to evaluate the OUT-PATSAT35 by modern psychometric analysis through IRT, in order to supplement the traditional CTT approach.

Methods

Design

The present study was performed using cross-sectional data from a multicenter, prospective cohort constituted of

determinants of patient satisfaction from ambulatory oncology and based on the OUT-PATSAT35 questionnaire [9]. This noninterventional, observational study was approved by the ethics committee of the Hospital of Besançon, the National French Data Protection Agency and supported by regional grant. Written informed consent was obtained from each patient.

Population and sample

Six hundred and ninety-two patients were recruited between January 2005 and December 2006 in two centers in eastern France (one university teaching hospital and one local hospital). Inclusion criteria were the following: age over 18 years, ability to understand written and spoken French, ability to provide written consent, ability to complete the questionnaire, confirmed histological diagnosis of cancer and ambulatory treatment by chemotherapy or radiotherapy due. Patients' socio-demographic and clinical characteristics were recorded at baseline.

Satisfaction with care assessment

Patients completed the OUT-PATSAT35 questionnaire at the beginning, at the end and 3 months after the end of the ambulatory treatment. Analyses presented in this paper were performed only on data collected at the end of the ambulatory treatment. At this time point, sample of respondents was large and diverse (several tumor locations) and the end of treatment was more relevant to assess satisfaction with care than at the beginning or 3 months after the end of the treatment.

The OUT-PATSAT35 questionnaire is organized into three sections: evaluating the medical and paramedical teams and the organization of the ambulatory department. It has two similar forms: the OUT-PATSAT35 RT and the OUT-PATSAT35 CT for patients who, respectively, receive ambulatory radiotherapy (RT) and chemotherapy (CT). It contains 35 items covering 12 multi-item scales and describes satisfaction with care in three domains of four scales each. The first two domains evaluate the doctors and the nurses (for chemotherapy) or the radiation therapists (for radiotherapy) in regard to their technical skills, their interpersonal skills, their ability in providing information and their availability, and the third domain evaluates the services and the care organization in regard to the exchange and to the quality of provided information, the waiting times and the physical environment. The last item is an overall satisfaction scale. The structure of the OUT-PATSAT35 is provided in Table 1. A five-level Likert response scale is used with the following response categories: "poor," "fair," "good," very good" or "excellent." All scores are linearly transformed to a 0–100 scale, with a higher score reflecting a higher level of satisfaction. When at least half of the item scores in a scale were missing, the patient score for that scale was missing. Otherwise, the score was equal to the mean of items answered. The calculation of the scores of the domains is similar.

Statistical analysis

CTT analysis

Questionnaire's acceptability was assessed via the patient's involvement in the study (>0.8 criteria) [10] and by the rates of the missing items and the complete questionnaires. The floor and ceiling effects on the scales and the domains were also computed. Convergent and divergent validity was evaluated using multitrait scaling analysis [11] conducted separately for the twelve scales and for the three domains: the doctors, the nurses or the radiation therapists and the services/care organization. The convergent validity of each item was assessed using the Spearman correlation's coefficient between each item and its own scale score, computed without including the corresponding item. The convergent validity was considered satisfactory if the correlation coefficient was higher than 0.40. For the divergent validity, the correlation between each item and its own scale score was expected to be greater than the correlation between the item and the other scale scores. Similar analyses were conducted on each domain.

The internal consistency reliability of the scales and the domains was assessed using the Cronbach's alpha coefficient [12] from available case analysis. It was expected to be higher than 0.70.

An exploratory factor analysis (EFA) was performed, and the factor structure was rotated using orthogonal rotations (varimax). Multiple criteria were examined such as the Kaiser–Guttman eigenvalues higher than 1.0 rule, the ratio of first to second eigenvalues, the variance explained and the interpretability of resulting factors [13]. This EFA was conducted to assess dimensionality of the questionnaire.

IRT analysis

CTT is based on the true score model (true score plus error). The major limitation of CTT is that the person ability (location on the latent variable) and the item difficulty, which influences the probability of a particular item response, cannot be estimated separately. Furthermore, the precision of measurement according to CTT depends on the ability level of the studied population. IRT offers advantages over CTT in assessing self-reported health outcomes. In contrast to CTT in which the respondents' observed scores on a whole questionnaire or scales are the

Table 1 Structure of the OUTPAT-SAT35 questionnaire

Items	Scales	Item content
Doctors' domain		
Item 1	Technical skills	Disease knowledge and experience *(connaissance/experience de la maladie)*
Item 2		Treatment and follow-up *(traitement et suivi médical)*
Item 3		Care for physical problems *(attention accordée aux problèmes physiques)*
Item 4	Interpersonal skills	Availability in listening to worries *(disponibilité à écouter les préoccupations)*
Item 5		Interest to the patient *(intérêt porté à la personne)*
Item 6		Comfort and support *(réconfort et soutien)*
Item 7	Information provided	Information about the disease *(informations fournies sur la maladie)*
Item 8		Information about medical examinations *(information fournies sur les examens médicaux)*
Item 9		Information about treatments *(informations fournies sur les traitements)*
Item 10	Availability	Promptness *(ponctualité à la consultation)*
Item 11		Time allowed *(temps consacré durant leur consultation)*
Nurses' or radiation therapists' domain		
Item 13	Technical skills	Treatment implementation *(manière dont ils ont pratiqué le traitement)*
Item 14		Care for physical comfort *(attention accordée au confort physique)*
Item 12	Interpersonal skills	Reception *(l' accueil pour le traitement)*
Item 15		Interest to the patient *(intérêt porté à la personne)*
Item 16		Comfort and support *(réconfort et soutien)*
Item 17		Human qualities *(qualités humaines)*
Item 18	Information provided	Information about medical examinations *(information fournies sur les examens médicaux)*
Item 19		Information about cares *(information fournies sur les soins)*
Item 20		Information about treatment *(informations fournies sur les traitements)*
Item 21	Availability	Speed in answering to particular requests *(rapidité de réponse aux demandes particulières)*
Item 22		Time allowed *(temps consacré)*
Services/organization domain		
Item 23	Exchange of information	Ease in identifying the referring doctor *(facilité d'identifier le médecin responsable)*
Item 24		Information consistency between members of the team *(cohérence des informations entre les membres du personnel soignant)*
Item 25		Exchange of information with services outside from the hospital *(échange d'information avec les services extra-hospitaliers)*
Item 26	Information provided	Kindness and helpfulness of the non-medical staff *(gentillesse et serviabilité du personnel d'accueil, secretariat, agents de service…)*
Item 27		Information about the organization of medical examination, treatment/cares *(informations fournies sur l'organisation des examens, du traitement ou des soins)*
Item 28		Informations about other services *(informations fournies sur l'ensemble des services disponibles)*
Item 29	Waiting time	Ease in calling the unit *(facilité à joinder le service par telephone)*
Item 30		Delay to get a medical appointment *(délai d'attente pour obtenir un rendez vous medical)*
Item 31		Speed of treatments' and examinations' execution *(rapidité d'exécution des examens et traitements)*
Item 32	Environment	Accessibility *(accessibilité)*
Item 33		Ease in finding the different units *(facilité à s'orienter vers les différents services)*
Item 34		Hospital's environment *(environnement de l'établissement)*
Overall satisfaction with care		
Item 35		Overall quality of care *(Qualité des soins reçus, de manière générale)*

The item content is given in French (in italic font) and translated in English language (in normal font)

unit of focus, in IRT, the item itself is the unit of focus. This mathematical measurement model assumes the link between the items' responses and the subject's location on an unmeasured underlying ("latent") trait [8, 14].

The IRT model choice depends on several considerations: the instrument's dimensionality (unidimensional or multidimensional), the number (dichotomous or polytomous item) and types (ordinal, nominal) of item response categories and the complexity of the model represented by the number of item parameters to estimate (1 to 4 parameters) [8, 14, 15]. The Rasch-type models, originating in the work of George Rasch [16], are a family of parametric unidimensional IRT models, which incorporate fewer item parameters than the other models. They present the advantages of simplicity in the interpretation, parsimony and robust estimation techniques achievable with small sample sizes [17]. Two models are available for polytomous items, with rank-ordered response categories: the partial credit model (PCM) [18, 19] and the rating scale model (RSM) [20]. The principal difference between these two models is that the distance between the thresholds (probabilistic midpoint between two adjacent response categories) is constrained to be equal across items (RSM) or not (PCM). In statistical terms, the RSM is nested within the PCM. Using data from the assessment, the Akaike information criterion (AIC) can be used to determine which model is the most appropriate to apply.

Analyses were conducted separately for each OUTPATSAT35 domain: the doctors' domain, the nurses' or the radiation therapists' domain and the services/organization's domain. The guidelines set out by Tennant [21] were followed. The distribution across the response categories of each item was examined. This examination is major to IRT analysis as it is important to ensure that the data adequately cover the full set of response options. Indeed, if very few respondents in the sample endorsed a response category, it could be appropriate to collapse two or more response options into one [15].

To assess the polytomous scales, the examination of the category structure was achieved. This approach tested whether the category ordering of the polytomous items worked as expected for constructing measurement (ordered/disordered thresholds). The empirical category probability curves of each item category were inspected in order to identify poor items and response choices. These curves for an ideal item exhibit steep trace lines with one distinct maximum and exceed all other category curves in one interval of the latent trait [22]. The distributions of persons and items were represented on the same logit scale. The quality of measure's targeting was assessed by the comparison of the mean location score obtained for persons with that of the value of zero set for the items. When the mean location for the persons is around the mean values of the items (value of zero), the items are well targeted for people in the sample. When the mean value is positive, the sample as a whole is located at a higher level than the average difficulty of the items (too easy). Contrarily, a negative mean value indicates that the sample is located at a lower level (too hard) [21].

The model fit to the data was explored with global and individual item-fit statistics. The Chi-square tests for each item compared the difference between observed and expected values defined by the model and assessed the invariance of item hierarchy across the measured construct being. At the individual level, if the p values associated with the Chi-square tests are significant after adjusting for multiple testing using Bonferroni corrections (inferior to the alpha type one error of 5 % divided by the number of item in the corresponding domain) [23], then the item is suspected to misfit the model expectations. Furthermore, if the residual statistics (standardized values) were outside the range ± 2.5, they are suspected to misfit the model expectations. High positive residuals are of particular concern (few respondents identified with bizarre, unexpected response patterns), whereas high negative residuals indicate some redundancy in the data.

Internal consistency of the domains was estimated by a Person Separation Index (PSI). This index can be interpreted equivalently to the Cronbach's alpha, and a PSI higher than 0.7 was required [21].

Finally, consideration must be given to the fundamental assumptions underlying IRT models. The assumption of unidimensionality, also known as the fact that the responses to the items only depended on one latent trait to characterize the individuals [8, 15, 21], was assessed through the analysis of the residuals correlation matrix of the model, for each domain. The absence of any meaningful pattern in the residuals (only correlation coefficients smaller than 0.3, or only disseminated high values on these coefficients among the correlation matrix) will support the assumption of unidimensionality and thus the absence of a second latent variable. The assumption of local independence, identified as the fact that the response to one specific item may not be dependent on the response to another item, was detected through the examination of the correlations' pattern among the residuals. Residuals' correlations higher than 0.3 were inferred to indicate response dependency [24]. The assumption of monotonicity, also known as the fact that the response to the items increases with the level of the latent trait, was assessed using Loevinger's H coefficient for each item. It is generally admitted that Loevinger's H coefficient greater than 0.3 allows establishing monotonicity of the items [25].

Rasch models are suitable for handling datasets that contain missing values [14, 26]. Thus, no respondent was excluded due to missing data, as the use of the Rasch modeling procedures was possible in that case.

Sample size requirements

Nowadays, few publications on health outcomes assessment make substantive rules to determine prestudy sample size calculations for IRT and how to compute the test's power. There is a lack of research to support the computations of sample size. Often based on practical experience, most authors recommend hundreds of patients for the simplest models (such as Rasch family models) [27, 28]. Furthermore, they advised large and diverse sample of respondents to ensure stable and good parameter's estimates for items in IRT models with multiple response categories. It is important to achieve responses in all item responses options and that the respondents in the dataset cover the full set of response categories.

Moreover, to date, a strong debate about the most appropriate fit statistic to use, the range of the fit statistic to employ when evaluating a fit and how the fit statistic should be interpreted is reported in the literature [28–31]. Apart from this, in the assessment of fit statistic in Rasch family models, it has been reported that fit residual appears to be less sensitive to changes in sample size, compared to Chi-square statistic. Numerous studies have shown that fit statistics is sample dependent [28]. The Chi-square statistics increases with increases in sample size, and so the probability of identifying a misfit where none was identified, also increase with sample size, especially when it is beyond 200. One advantage of the Rasch-type models over traditional psychometric methods is to allow robust parameter estimates achievable with small sample size. However, in the current study, the large sample might alter global and item fit and considerably increased the power of the tests.

Among the 692 patients included in the cohort, 605 of them had completed the OUT-PAT35 questionnaire at the end of treatment. This sample size would allow a large power and is enough for the CTT analysis (175–350 patients can be considered as enough) [11] and IRT analysis [32]. Nevertheless, this sample size can create significant results for only small deviation of the assumed model. Consequently, a sensitivity analysis has been conducted by adjusting the statistics of the fit test on a virtual sample of 250 individuals assumed to be powerful enough to detect whether the Rasch model fit, or not the data ($N = 250$) [32].

Data were analyzed using SAS® (Version 9.2, SAS Institute Inc, Cary, NC) and RUMM2030 software [33]. All tests with multiple comparisons were two-sided, and the type I error was set to a Bonferroni-adjusted value (5 % divided by the number of item in the corresponding domain) [23].

Results

Study population

Six hundred and ninety-two patients were recruited between January 2005 and December 2006, and the characteristics of the included respondents were fully described elsewhere [9]. Six hundred and five patients (87.4 %) completed the OUT-PATSAT35 at the end of the ambulatory treatment which is the time of interest of the present study. The mean age of the patients was 64 years (range 29–88), and 52 % of them were men. Eighty-four percent (508) of the respondents had ambulatory RT.

Classical psychometric analysis according to classical test theory

The participation rate to the questionnaire at the end of the ambulatory treatment was 87.4 %. Of the 605 respondents at this time point, 352 questionnaires (58.2 %) were fully completed and the rate of item completion was 94.3 %. Scores were high for all OUT-PATSAT35 scales (mean 58.2–74.5 on a scale of 0–100) and all domains (mean 62.5–66.6 on a scale of 0–100). The ceiling effects varied between 6.7 and 27.6 % in the OUTPAT-SAT35 scales and between 3 and 11.1 % in the domains of the questionnaire. All floor effects were lower than 3 %. Cronbach's alpha coefficient was superior to 80 % in all domains and all scales, except for the "doctors availability" scale ($\alpha = 0.79$) and for the "environment" scale ($\alpha = 0.74$). All these results are presented in Table 2.

The EFA extracted a total of three factors with eigenvalues significantly greater than 1.0. These factors broadly corresponded to the main three domains of the questionnaire. They collectively represented almost 71 % of the variance with 31, 25 and 15 % of the variance, respectively, explained by each factor. The twelve scales of the questionnaire were not clearly identified by this analysis.

In the multitrait scaling analysis on the scales and the domains, the convergent validity exceeded the 0.4 criterion for all items. The item divergent validity criteria in the scales of each domain were not respected for eight items (items 3 and 11 in the doctors' scales—items 12–14 in the nurses' or radiation therapists' scales—items 25 and 26 in the services/organization scales) (Table 3). However, on the domain level, all items respected the divergent validity criteria.

Modern psychometric analysis according to IRT

Very few respondents in the sample endorsed the response category "poor," encoded one, for all items (1.8 %).

Table 2 Satisfaction score scales and domains according to OUT-PATSAT35 questionnaire, structure and internal consistency

Scales	Number and items	*N* (%)	Range	Mean (SD)	Cronbach's α	Floor effect (%)	Ceiling effect (%)
Doctors' domain	11 items	593 (98.0)	4.5–100	63.8 (21.2)	0.96	0	6.2
Doctors' technical skills	3 items—1 to 3	592 (97.9)	16.7–100	69.2 (20.5)	0.90	0	16.2
Doctors' interpersonal skills	3 items—4 to 6	588 (97.2)	0–100	64.3 (24.0)	0.95	0.9	15.1
Doctors' information provision	3 items—7 to 9	592 (97.9)	0–100	61.7 (25.2)	0.95	1.4	14.5
Doctors availability	2 items—10 and 11	595 (98.3)	0–100	58.4 (23.7)	0.79	1.0	9.6
Nurses' or radiation therapists' domain	11 items	597 (98.7)	11.4–100	66.7 (20.5)	0.97	0	11.1
Nurses' technical skills	2 items—13 and 14	601 (99.3)	0–100	72.5 (20.3)	0.90	0.3	22.5
Nurses' interpersonal skills	4 items—12, 15 to 17	599 (99.0)	6.3–100	71.1 (20.6)	0.94	0	17.4
Nurses' information provision	3 items—18 to 20	573 (94.7)	0–100	58.2 (25.1)	0.98	2.6	11.9
Nurses' availability	2 items—21 and 22	590 (97.5)	0–100	64.6 (23.0)	0.90	0.3	16.4
Services' and care organization's domain	12 items	588 (97.2)	12.5–100	62.7 (17.5)	0.92	0	3.0
Exchange of information	3 items—23 to 25	574 (94.9)	0–100	64.9 (21.3)	0.86	0.2	12.7
Information provided	3 items—26 to 28	569 (94.0)	8.3–100	63.6 (21.5)	0.83	0	11.6
Waiting time	3 items—29 to 31	537 (88.8)	0–100	61.2 (19.5)	0.84	0.2	6.9
Environment	3 items—32 to 34	584 (96.5)	8.3–100	61.0 (20.0)	0.74	0	6.7
Overall satisfaction	1 item—35	590 (97.5)	25–100	74.5 (19.5)		0	27.6

N: number of respondents; mean ± standard deviations (SDs) of the scores in the OUT-PATSAT35 questionnaire; range is the range of the scores in each scale or domain; scores in all scales or domains range from 0 to 100, with higher scores representing greater levels of satisfaction

Therefore, the two lowest response scale values were combined into one. We compared the classical psychometric properties for the original version of the questionnaire (5-category) and the 4-category version. The two questionnaire versions gave similar results, and we thus decided to perform the IRT analyses on the 4-category version of the questionnaire.

Using data from the assessments, the AIC confirmed that a PCM was more appropriate than a RSM. The threshold pattern for each item of each domain did not show the presence of disordered threshold for the items. The category response pattern of each domain is shown in Figure A in the supplementary files. Visual inspection of the category probability curves for all items showed an appropriate rank order for all categories (data not shown). The distributions of persons and items on the same logit scale are shown for each domain in Figure B in the supplementary files. The mean ± SD of the person estimate was 0.321 ± 2.616 in the doctors' domain, 0.838 ± 3.423 in the nurses' or radiation therapists' domain and 0.134 ± 1.668 in the services/organization domain. Therefore, with scales centered on zero value, these scales seemed to be well targeted for this sample.

The assumption of monotonicity, assessed using Loevinger's H coefficient (Table 4), was respected on each domain. The analysis of the residuals showed a major deviation from unidimensionality only on the nurses' or radiation therapists' domain. The residual correlation matrix is shown in Table A in supplementary files.

The doctors' domain

The upper part of Table 4 shows fit of the model for the doctors' domain. The total-item Chi-square in this domain was 364.69, and the *p* value was inferior to 0.001. All the *p* values adjusted by Bonferroni correction on the item Chi-square statistics were not significant except for the item 10 about "promptness" (*p* value <0.001). Eight items (items 2, 3, 4, 5, 6, 7, 8 and 10) had residuals fit indices outside the range ±2.5, but only the item 10 presented high and positive values for this index. For the sensitivity analysis performed on a virtual sample of 250 patients, the overall model fit was slightly improved (Chi-square = 164.87, *p* value <0.001). The upper part of Table 5 shows fit of the model on the total sample and on a virtual sample of 250 respondents without the item 10 in the doctors' domain. Dropping the item 10 (significant Chi-square statistics and high positive residual fit value, outside the range ±2.5) from the item pool and performing the fit test on a virtual sample of 250 patients enhance the overall item fit (Chi-square = 59.66, *p* value = 0.99).

Internal consistency was high with a PSI of 0.93. A local dependence was found between different pairs of items including all the items 1–9. Highest positive residual correlations were detected between items 7–9 which belong to "information provided" scale. The residual correlation matrix is shown in Table A in supplementary files.

Table 3 Multitrait analysis: correlation of each item with their own scale score, other scale scores, their own domain score and other domain scores

Item	Scales in each domain				Doctors' domain	Nurses' or radiation therapists' domain	Service's and care organization's domain
	Technical skills	Interpersonal skills	Information provided	Availability			
Doctors' domain							
Item 1	**0.798**	0.708	0.671	0.579	**0.754**	0.592	0.628
Item 2	**0.876**	0.785	0.729	0.634	**0.830**	0.622	0.674
Item 3	***0.772***	0.876	0.726	0.647	**0.846**	0.642	0.668
Item 4	0.843	**0.879**	0.771	0.685	**0.876**	0.657	0.700
Item 5	0.835	**0.903**	0.750	0.673	**0.863**	0.671	0.695
Item 6	0.804	**0.877**	0.763	0.674	**0.862**	0.674	0.697
Item 7	0.749	0.769	**0.899**	0.641	**0.845**	0.649	0.667
Item 8	0.751	0.765	**0.907**	0.654	**0.848**	0.667	0.680
Item 9	0.711	0.735	**0.883**	0.649	**0.819**	0.648	0.682
Item 10	0.535	0.542	0.532	**0.647**	**0.589**	0.482	0.565
Item 11	0.701	0.739	0.696	***0.647***	**0.781**	0.604	0.663
Nurses' or radiation therapists' domain							
Item 12	0.880	***0.834***	0.644	0.748	0.635	**0.823**	0.692
Item 13	***0.834***	0.842	0.621	0.722	0.603	**0.805**	0.684
Item 14	***0.834***	0.884	0.661	0.754	0.613	**0.844**	0.665
Item 15	0.878	**0.894**	0.745	0.812	0.647	**0.889**	0.703
Item 16	0.805	**0.862**	0.776	0.806	0.649	**0.874**	0.671
Item 17	0.803	**0.813**	0.642	0.806	0.570	**0.797**	0.668
Item 18	0.651	0.738	**0.941**	0.792	0.691	**0.834**	0.695
Item 19	0.673	0.756	**0.964**	0.800	0.688	**0.848**	0.686
Item 20	0.680	0.756	**0.950**	0.804	0.692	**0.850**	0.701
Item 21	0.745	0.791	0.783	**0.820**	0.669	**0.848**	0.713
Item 22	0.752	0.810	0.762	**0.820**	0.673	**0.852**	0.719

	Exchange of information	Information provided	Waiting time	Environment			
Service's and care organization's domain							
Item 23	**0.703**	0.639	0.602	0.440	0.687	0.572	**0.686**
Item 24	**0.767**	0.704	0.611	0.472	0.687	0.689	**0.754**
Item 25	***0.722***	0.767	0.609	0.482	0.688	0.672	**0.745**
Item 26	0.635	***0.605***	0.579	0.452	0.595	0.681	**0.679**
Item 27	0.753	**0.771**	0.623	0.467	0.731	0.765	**0.784**
Item 28	0.651	**0.715**	0.533	0.414	0.643	0.656	**0.669**
Item 29	0.575	0.547	**0.637**	0.577	0.523	0.533	**0.661**
Item 30	0.606	0.606	**0.746**	0.495	0.605	0.575	**0.688**
Item 31	0.642	0.622	**0.696**	0.517	0.606	0.670	**0.713**
Item 32	0.294	0.300	0.384	**0.525**	0.305	0.302	**0.422**
Item 33	0.524	0.537	0.590	**0.599**	0.478	0.490	**0.649**
Item 34	0.447	0.492	0.491	**0.514**	0.454	0.433	**0.547**

Multitrait analysis in the three domains of the questionnaire: scales of the doctors' domain, scales of the nurses' or radiation therapists' domain and scales of the service/organization' domain. Correlations between item own scale and item other scales of the domain are presented in each domain. Correlations between item own domain and other domain are also presented. Values in bold and in italic bold font are the item own scale correlation and the item own domain correlation (convergent validity). Numbers in bold font only are item own scale correlation higher than item correlation with the other scales of the domain and item own domain correlation higher than item correlation with the other domain (divergent validity). Numbers in normal font are correlations between the items and the other scales of the domain and between the items and other domain

Table 4 Difficulties, Loevinger's H coefficient and fit statistics for the OUT-PATSAT35 items and domains

Item	Item difficult (logit)	SE	Fit residual	Total sample ($n = 605$)		Virtual sample ($n = 250$)		Loevinger's H coefficient
				χ^2	P value	χ^2	P value	
Doctors' domain				364.69	**<0.001**	164.87	**<0.001**	0.74
1	−1.622	0.083	1.414	11.89	0.2197	5.37	0.80	0.74
2	−0.863	0.081	***−3.303***	19.93	0.0183	9.01	0.436	0.79
3	−0.035	0.076	***−4.113***	13.42	0.1446	6.07	0.733	0.77
4	−0.027	0.077	***−6.213***	20.50	0.0151	9.27	0.413	0.79
5	−0.078	0.075	***−5.546***	17.18	0.0460	7.77	0.558	0.78
6	0.184	0.076	***−4.158***	14.32	0.1113	6.48	0.692	0.77
7	0.183	0.074	***−2.688***	13.22	0.1530	5.98	0.742	0.75
8	0.519	0.074	***−2.771***	11.26	0.2582	5.09	0.826	0.77
9	0.312	0.073	−1.100	6.36	0.7032	2.88	0.969	0.74
10	0.974	0.072	***10.472***	220.43	**<0.0001**	99.66	**<0.001**	0.58
11	0.352	0.074	1.575	16.18	0.0632	7.32	0.604	0.71
Nurses' or radiation therapists' domain				304.16	**<0.001**	142.93	**0.002**	0.83
12	−1.498	0.092	0.298	33.32	**0.0001**	15.66	0.074	0.82
13	−1.603	0.091	0.066	37.32	**<0.0001**	17.54	0.041	0.81
14	−0.782	0.087	−2.167	28.32	**0.0008**	13.31	0.149	0.84
15	−0.198	0.087	***−5.769***	22.64	0.0070	10.64	0.301	0.87
16	0.301	0.087	***−2.847***	11.28	0.2568	5.30	0.807	0.84
17	−1.838	0.089	−0.001	42.75	**<0.0001**	20.09	0.017	0.81
18	1.663	0.091	0.379	37.12	**<0.0001**	17.44	0.042	0.82
19	1.707	0.089	−0.682	34.71	**<0.0001**	16.31	0.061	0.83
20	1.471	0.087	−0.691	42.24	**<0.0001**	19.85	0.019	0.83
21	0.396	0.087	0.267	10.19	0.3353	4.79	0.852	0.82
22	0.382	0.086	−0.241	4.27	0.8931	2.00	0.991	0.83
Services/organization domain				281.88	**<0.001**	120.88	0.187	0.56
23	−0.593	0.064	−0.561	16.64	0.0546	7.14	0.623	0.57
24	−0.158	0.066	***−3.327***	18.94	0.0257	8.12	0.522	0.61
25	0.288	0.072	***−2.990***	22.69	0.0069	9.12	0.373	0.62
26	−1.109	0.067	−1.944	10.44	0.3158	4.48	0.877	0.59
27	0.171	0.065	***−4.062***	31.54	**0.0002**	13.52	0.140	0.62
28	0.655	0.068	−0.300	5.12	0.8242	2.19	0.988	0.58
29	−0.068	0.070	−0.361	2.95	0.9661	1.27	0.999	0.56
30	0.523	0.073	−0.707	7.95	0.5396	3.41	0.945	0.59
31	0.038	0.068	−1.671	9.12	0.4262	3.91	0.917	0.58
32	0.668	0.061	***8.959***	128.75	**<0.0001**	55.21	**<0.001**	0.39
33	0.069	0.067	−0.361	9.08	0.4302	3.89	0.918	0.54
34	−0.484	0.063	***3.752***	18.66	0.0282	8.00	0.533	0.47

When the fit residual was outside range ±2.5 (italic bold font) and/or the Chi-square test was significant (bold font), the misfit was suspected

The nurses' or radiation therapists' domain

The middle part of Table 4 shows fit of the model for the nurses or radiation therapists' domain. The total-item Chi-square in this domain was 304.16 (p value <0.001). Bonferroni corrected p values of the item Chi-square statistics were significant for items 12–14 and 17–20. All items had residual fit indices inside the range ±2.5 except items 15 and 16. Regarding the sensitivity analysis performed on a virtual sample of 250 patients, the overall model fit was slightly improved (Chi-square = 142.93, p value = 0.002).

Table 5 Difficulties and fit statistics for the OUT-PATSAT35 items and domains (without the item 10 for the doctors' domain and without the items 32 and 34 for the services/organization's domain)

Item	Item difficult (logit)	SE	Fit residual	Total sample ($n = 605$)		Virtual sample ($n = 250$)	
				χ^2	P value	χ^2	P value
Doctors' domain				127.43	**0.006**	59.66	0.994
1	−1.612	0.087	3.040	22.45	0.007	10.51	0.311
2	−0.817	0.085	−1.927	9.25	0.414	4.33	0.888
3	−0.058	0.080	***−3.266***	9.88	0.360	4.63	0.866
4	−0.071	0.081	***−5.640***	14.53	0.105	6.80	0.657
5	−0.008	0.080	***−7.777***	14.08	0.120	6.59	0.679
6	0.303	0.080	***−3.152***	9.04	0.434	4.23	0.896
7	0.400	0.078	−1.772	10.03	0.348	4.70	0.859
8	0.669	0.078	−1.635	5.04	0.831	2.36	0.984
9	0.439	0.078	0.421	5.78	0.762	2.71	0.975
11	0.487	0.078	***4.594***	27.36	**0.001**	12.81	0.171
Services/organization domain				132.46	**0.002**	57.80	0.996
23	−0.650	0.068	0.847	11.15	0.266	4.86	0.846
24	−0.159	0.071	***−2.862***	13.77	0.131	6.01	0.739
25	0.349	0.077	***−2.778***	16.49	0.057	7.20	0.617
26	−1.198	0.071	−0.699	11.33	0.254	4.94	0.839
27	0.201	0.069	***−3.974***	23.55	0.005	10.27	0.329
28	0.763	0.073	0.695	5.12	0.823	2.24	0.987
29	−0.053	0.074	***2.290***	8.21	0.513	3.58	0.937
30	0.604	0.077	0.037	11.01	0.275	4.81	0.851
31	0.059	0.072	−0.510	6.29	0.711	2.74	0.974
33	0.084	0.084	***3.702***	25.55	**0.002**	11.15	0.267

When the fit residual was outside range ±2.5 (italic bold font) and/or the Chi-square test was significant (bold font), the misfit was suspected

Internal consistency of the domain was equal to 0.95. Most of the local dependence was found between several items of the domain and items 18–20 which belong to "information provided" scale. Highest positive residuals correlations were detected between them (items 18–20). The residual correlation matrix is shown in Table A in supplementary files.

The services/organization domain

The lower part of Table 4 shows fit of the model for the services/organization domain. The total-item Chi-square in this domain was 281.88, and the p value was inferior to 0.001. Bonferroni corrected p values of the item Chi-square statistics were significant for two items 27 and 32, and five items had residual fit indices outside the range ±2.5 (items 32 and 34 with positive values and items 24, 25 and 27 with negative values). For the sensitivity analysis performed on a virtual sample of 250 patients, the overall model fit was good (Chi-square = 120.88, p value = 0.19). The lower part of Table 5 shows fit of the model on the total sample and on a virtual sample without the items 32 and 34 about accessibility and hospital's environment, respectively. Dropping the items 32 and 34 (most significant Chi-square statistics and/or high positive residual fit index, outside the range ±2.5) from the item pool and performing the analysis on a virtual sample enhance the overall item fit (Chi-square = 90.71, p value = 0.71).

Internal consistency was high with a PSI of 0.90. A local dependence was found between items 32 and 24 ($r = -0.33$), items 25 and 34 ($r = -0.36$), items 27 and 28 ($r = 0.36$), and items 32 and 27 ($r = -0.39$). The residual correlation matrix is shown in Table A in supplementary files.

Discussion

In this study, we presented the results of evaluation of the psychometric properties of the French OUT-PATSAT35 questionnaire using two supplementary analyses: CTT and IRT. A large and heterogeneous sample of patients and few missing data made this study powerful.

According to CTT analyses, the main psychometric properties of OUT-PATSAT35 questionnaire were confirmed. A good acceptability of the questionnaire and a good internal consistency (as measured by Cronbach's alpha) were observed in most scales. The item convergent validity was satisfying, but a poor divergent validity was reported for several items in each domain, according to their own scales: items 3 and 11 in the doctors' domain, items 12–14 in the nurses or radiation therapists' domain and items 25 and 26 in the services/organization domain. However, on the domain level, all items respected the divergent validity criteria. Similar results were found in the previous studies which validated the Spanish and French versions of this questionnaire and which were conducted on smaller samples [5–7]. Otherwise, the scores of the OUT-PATSAT35 questionnaire indicated that the levels of SC were overall good, even though a very low floor effect was found. Indeed, a very small number (<2 %) of respondents in the present study chose the response category associated with the lowest level of satisfaction with care. Similar results were reported in the previous Spanish studies [6, 7]. Consequently, it would be more appropriate to develop a 4-category version of the questionnaire. However, it is well accepted that patients may be reluctant to choose the lowest and worse response category, and so, reducing the number of response options would only shift the problem. As an alternative, it would be interesting to revise the response categories' denominations or to adapt the scoring according to this consideration (to combine the two lowest categories into one just for the scoring).

The IRT analysis of each dimension of the OUT-PATSAT35 showed that the unrestricted PCM was the most appropriate model. The threshold distances varied across items. The IRT analysis was performed on a 4-category version of the questionnaire. The inspection of the threshold patterns and the option characteristic curves for each item revealed a right rank order for all categories. Thus, the response choices of the items discriminate well. The quality of targeting of persons and items was good; thus, the items of the French OUT-PATSAT35 questionnaire were well targeted for people in this sample of patients in ambulatory oncology.

Each domain showed a lack of fit of the model, with several items assumed to misfit.

In the doctors' domain, the item 10 presented the highest fitting default with the model. Moreover, high positive residual characterized this item. We assume that this result is due to more aberrant response patterns than to redundancy. This item assesses doctors' promptness, and one may wonder whether it is important in the assessment of satisfaction with doctors' care. A misunderstanding of the item can also be discussed. Dropping this item from the item pool and performing the analysis on a smaller virtual sample allows obtaining a correct fit of the Rasch model.

In the nurses' or radiation therapists' domain, nine items out of eleven in this domain were suspected to misfit. A high negative residual was found for the item 15 (interest of the healthcare team) and the item 16 (comfort and support), which suggested some redundancy. Chi-square statistics were slightly enhanced when it was adjusted on the virtual sample.

In the services/organization domain, the items 32 and 34 were concerned with the highest default of fit of the model. Indeed, for these items, a high positive residual was found. The content of these items can be questioned as to what extent the accessibility to hospital and its environment are important among patients to assess SC. An analysis without these items and on a smaller virtual sample enhanced all the fit statistics (global and individual), which became correct.

In the three domains, the assumption of local independency was not respected. A high positive local dependency was particularly detected between items about "information provided" in each domain: the items 7–9 in the doctor's domain, the items 18–20 in the nurses' or radiation therapists' domain and the items 27 and 28 in the services/organization domain. Consequently, it would be appropriate to collapse these "information" items into one, in each domain, or to create a fourth domain about information provided which contains all these items. Moreover, most of the items concerned by the local dependence presented a negative residuals fit which could indicate some redundancy in the data. This consideration was found in each domain except in the nurses' or radiation therapists' domain where the analysis suggested a major default of unidimensionality. Thus, the data in this domain seem to describe more than a single underlying construct and further work is required to determine whether this domain is best represented by two or more latent variables. Consequently, some caution needs to be exercised in interpreting the scores from this domain and the changes in this score.

Our study highlights the complementary relationship between the IRT and CTT analyses. The CTT and the IRT analyses seemed to provide different results and to suggest some discrepancies. Indeed, the CTT analyses brought out problems with the items 3, 11–14, 25 and 26, whereas the IRT analyses identified different misfitting items: the items 10, 32 and 34, in particular. Comparisons between the results of the CTT and the Rasch analysis should be carried out with caution as each method differs on several aspects of measurement. First, in contrast to CTT in which the scores (on scales or domains) are the unit of focus, in IRT, the item itself is the unit of focus. Second, the strength of the IRT models is that the results of each item are independent of the sample and of the other items in the

questionnaire [8, 14, 19]. Finally, in our study in particular, the two analyses were not performed at the same level of the questionnaire's construct. The CTT analyses were performed at the scales' level as well as at the domains' level, whereas the IRT analyses were only performed at the domains' level. The problems identified by the CTT analyses concerned the scales structure of the questionnaire. Similar CTT analyses at the domains' level did not find default in the questionnaire's construct. Nevertheless, the items misfit observed from the Rasch analysis (items 10, 32 and 34) broadly corresponded to anomalies observed from the multitrait scaling analysis conducted for the three domains, in CTT. Indeed, a more thorough examination brought to light the items 10, 32 and 34, which presented the smallest correlation coefficients between items with their own domain (with an important difference with the correlation coefficients of the other items). These misfits, slightly present in the CTT, were identified more clearly in the IRT analyses. Thus, the results of the CTT analysis and the IRT analysis of the OUT-PATSAT35 mostly agreed. A good internal consistency of the domains was found by the two analyses.

Moreover, the results of the dimensionality or factor structure assessment, using both CTT and IRT analyses, differ. However, caution should be also exercised when comparing the results of the EFA in CTT analysis and the results of the residuals analysis in IRT. Indeed, the aim of the EFA computed on a whole questionnaire is to assess dimensionality of the whole questionnaire (to identify factors within a correlation matrix), whereas for IRT analysis computed on each domain, the aim is to identify whether multidimensionality exists in the residuals once the unidimensional structure has been removed in each domain (and not the whole questionnaire) and whether the responses to the item pool can be explained by a multidimensional latent trait [34, 35]. On another point, the IRT identified more clearly problems in the structure of the OUT-PATSAT35 questionnaire where none was identified by the CTT, even in the previous studies, such as the redundancy between items, problems in the response category and misfitting items. Thus, the IRT offers many advantages over CTT, mostly in the development, refinement and evaluation of a questionnaire or a reduced form, but without replacing the CTT.

The findings of this study indicate that a refinement of the questionnaire could be necessary. One may wonder how misfitting items should be treated. Removing items from the domains, which do not fit the IRT model, is advocated and often may not have impact on the measurement properties of the questionnaire or improve it [34, 36, 37]. Furthermore, studies in the Rasch literature have investigated the impact on clinical utility of the questionnaire to remove these misfitting items (impact on item locations, on the ability of the revised questionnaire to detect significant changes in scores between different patient groups in clinical trial) [34, 38]. No significant impact was found, but caution was advised.

To conclude, the OUTPAT-SAT35 questionnaire does not meet expectations of the IRT measurement model. The Rasch analysis revealed misfitting and redundant items. Taking the above problems into consideration, it could be interesting to refine the questionnaire in a future study.

Acknowledgments The authors thank all physicians from the centers participating in the study who agreed to invite patients to participate in this study. We thank the clinical research assistants in the two centers who participated in the data collection. This work was supported by the Regional French Hospital Clinical Research Program.

Conflict of interest The authors have no potential conflict of interest.

References

1. Burke, L. (2006). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims: draft guidance. *Health and Quality of Life Outcomes, 4*, 1.
2. Rubin, H. R., Gandek, B., Rogers, W. H., Kosinski, M., McHorney, C. A., & Ware, J. E., Jr. (1993). Patients' ratings of outpatient visits in different practice settings. Results from the Medical Outcomes Study. *JAMA, 270*(7), 835–840.
3. Borras, J. M., Sanchez-Hernandez, A., Navarro, M., Martinez, M., Mendez, E., Ponton, J. L., et al. (2001). Compliance, satisfaction, and quality of life of patients with colorectal cancer receiving home chemotherapy or outpatient treatment: A randomised controlled trial. *BMJ, 322*(7290), 826.
4. Bredart, A., Bottomley, A., Blazeby, J. M., Conroy, T., Coens, C., D'Haese, S., et al. (2005). An international prospective study of the EORTC cancer in-patient satisfaction with care measure (EORTC IN-PATSAT32). *European Journal of Cancer, 41*(14), 2120–2131.
5. Poinsot, R., Altmeyer, A., Conroy, T., Savignoni, A., Asselain, B., Leonard, I., et al. (2006). Multisite validation study of questionnaire assessing out-patient satisfaction with care questionnaire in ambulatory chemotherapy or radiotherapy treatment. *Bulletin du Cancer, 93*(3), 315–327.
6. Arraras, J. I., Illarramendi, J. J., Viudez, A., Lecumberri, M. J., de la Cruz, S., Hernandez, B., et al. (2012). The cancer outpatient satisfaction with care questionnaire for chemotherapy, OUT-PATSAT35 CT: A validation study for Spanish patients. *Supportive Care in Cancer, 20*(12), 3269–3278.
7. Arraras, J. I., Rico, M., Vila, M., Chicata, V., Asin, G., Martinez, M., et al. (2010). The EORTC cancer outpatient satisfaction with care questionnaire in ambulatory radiotherapy: EORTC OUT-PATSAT35 RT. Validation study for Spanish patients. *Psycho-oncology, 19*(6), 657–664.
8. Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9 Suppl), II28–II42.
9. Nguyen, T. V., Bosset, J. F., Monnier, A., Fournier, J., Perrin, V., Baumann, C., et al. (2011). Determinants of patient satisfaction in ambulatory oncology: A cross sectional study based on the OUT-PATSAT35 questionnaire. *BMC Cancer, 11*, 526.

10. Sitzia, J., & Wood, N. (1998). Response rate in patient satisfaction research: an analysis of 210 published studies. *International Journal for Quality in Health Care, 10*(4), 311–317.
11. Nunnally, J. C. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.
12. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.
13. Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286.
14. De Ayala, R. (2009). *The theory and practice of item response theory*. New York: Guilford Press.
15. Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(Suppl 1), 5–18.
16. Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* ERIC.
17. Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. Berlin: Springer.
18. Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.
19. Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.
20. Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573.
21. Tennant, A., & Conaghan, P. G. (2007). 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 and Rheumatism, 57*(8), 1358–1362.
22. Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the headache impact test (HIT). *Quality of Life Research, 12*(8), 913–933.
23. Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ, 310*(6973), 170.
24. Ramp, M., Khan, F., Misajon, R. A., & Pallant, J. F. (2009). Rasch analysis of the Multiple Sclerosis Impact Scale MSIS-29. *Health Qual Life Outcomes, 7*, 58.
25. Sijtsma, K., & Molenaar, I. W. (Eds.). (2002). *Introduction to nonparametric item response theory* (Vol. 5). Beverly Hills: Sage.
26. Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. London: Psychology Press.
27. Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.
28. 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.
29. 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.
30. Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*(2), 152–176.
31. Linacre, J., & Wright, B. (1994). Chi square fit statistics. *Rasch Measurement Transactions, 8*(2), 350.
32. Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Standardized chi square fit statistic. *Rasch Measurement Transactions, 17*, 918.
33. Sheridan, B. (1998). RUMM item analysis package: Rasch unidimensional measurement model. *Rasch Measurement Transactions, 11*(4), 599.
34. Smith, A. B., Wright, P., Selby, P. J., & Velikova, G. (2007). A Rasch and factor analysis of the Functional Assessment of Cancer Therapy-General (FACT-G). *Health Qual Life Outcomes, 5*, 19.
35. Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*(2), 205–231.
36. Petersen, M. A., Groenvold, M., Aaronson, N., Blazeby, J., Brandberg, Y., de Graeff, A., et al. (2006). Item response theory was used to shorten EORTC QLQ-C30 scales for use in palliative care. *Journal of Clinical Epidemiology, 59*(1), 36–44.
37. Pallant, J. F., Miller, R. L., & Tennant, A. (2006). Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. *BMC Psychiatry, 6*, 28.
38. Smith, A. B., Wright, E. P., Rush, R., Stark, D. P., Velikova, G., & Selby, P. J. (2006). Rasch analysis of the dimensional structure of the Hospital Anxiety and Depression Scale. *Psychooncology, 15*(9), 817–827.