Brief Methodological Report

Rasch Analysis of the Edmonton Symptom Assessment System

Emma Sprague, MHSc(Nursing), Richard J. Siegert, PhD, Oleg Medvedev, PhD, and Margaret H. Roberts, BHSc(Nursing), DClinPsych

Hospice West Auckland (E.S.), Auckland; School of Clinical Sciences (R.J.S., M.H.R.), Auckland University of Technology (AUT), Auckland; and Centre for Medical and Health Sciences Education (O.M.), School of Medicine, University of Auckland, Auckland, New Zealand

Abstract

Context. The Edmonton Symptom Assessment System (ESAS) is a widely used multisymptom assessment tool in cancer and palliative care settings, but its psychometric properties have not been widely tested using modern psychometric methods such as Rasch analysis.

Objectives. To apply Rasch analysis to the ESAS in a community palliative care setting and determine its suitability for assessing symptom burden in this group.

Methods. ESAS data collected from 229 patients enrolled in a community hospice service were evaluated using a partial credit Rasch model with RUMM2030 software (RUMM Laboratory Pty, Ltd., Duncraig, WA). Where disordered thresholds were discovered, item rescoring was undertaken. Rasch model fit and differential item functioning were evaluated after each iterative phase.

Results. Uniform rescoring was necessary for all 12 items to display ordered thresholds. The best model fit was achieved after item rescoring and combining three pairs of locally dependent items into three superitems ($\chi^2 = 29.56$ [27]; $P = 0.33$) that permitted ordinal-to-interval conversion.

Conclusion. The ESAS satisfied unidimensional Rasch model expectations in a 12-item format after minor modifications. This included uniform rescoring of the disordered response categories and creating superitems to improve model fit and clinical utility. The accuracy of the ESAS scores can be improved by using ordinal-to-interval conversion tables published in the article. J Pain Symptom Manage 2018;55:1356–1363. © *2018 The Authors. Published by Elsevier Inc. on behalf of American Academy of Hospice and Palliative Medicine. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Key Words

Edmonton Symptom Assessment System (ESAS), Rasch analysis, psychometrics, outcome measurement, palliative care

Introduction

The assessment of symptom burden is one of the cornerstones of effective treatment within palliative care. The Edmonton Symptom Assessment System (ESAS)[1] is the most common multisymptom assessment measure used internationally throughout cancer and palliative care settings.[2] Despite widespread clinical use, the ESAS[1] has undergone limited psychometric evaluation. Consequently, it is unclear whether the patient's ESAS[1] scores are an accurate reflection of symptom burden or are impacted by measurement error.

Rasch analysis is a means of psychometrically evaluating measurement tools that offers a number of advantages for clinical measures such as the ESAS,[1] when compared with more commonplace methods such as Classical Test Theory.[3] These advantages include the construction of an interval scale, where

Address correspondence to: Margaret H. Roberts, BHSc(Nursing), DClinPsych, School of Clinical Sciences, Auckland University of Technology (AUT), Auckland, New Zealand. E-mail: Margaret.roberts@aut.ac.nz

Accepted for publication: January 26, 2018.

0885-3924/$ - see front matter
https://doi.org/10.1016/j.jpainsymman.2018.01.016

the distance between response options given to the test taker is equal. Rasch analysis enables the evaluation of the appropriate stochastic ordering of items, freedom from item bias, and in the case of polytomous items—evaluating the appropriate ordering of responses.[3,4] Local independence assumptions are assessed, including unidimensionality, requiring all items to be functionally dependent on only one underlying dimension (e.g., symptom burden). If the data fit the Rasch model, then patients can be ordered in terms of their ability, and the items ordered in terms of their difficulty, using the same interval-level scale.[5] In the case of the ESAS,[1] ability refers to the level of symptom burden the patient has, and difficulty refers to the level of symptom burden implied by an item. For example, an item about shortness of breath, indicating a high level of symptom burden, would be a very difficult item. This can be presented graphically in a person-item threshold distribution plot, which enables the researcher to view how well the range of item difficulty relates to the range of sample abilities. Furthermore, this process enables a visual account of floor and ceiling effects.[6,7]

Rasch analysis provides the ability to evaluate item bias or differential item functioning (DIF) based on the characteristics of the individual test takers. DIF occurs when people from different groups (e.g., males and females) with the same ability on the latent trait being measured score differently on an item. In the case of the ESAS,[1] we may wish to ensure that patients are responding on the basis of individual differences in their level of shortness of breath, independent on their demographic characteristics (e.g., ethnicity). Rasch analysis enables researchers to evaluate and rescore items displaying disordered response categories to improve its measurement properties and clinical utility.[3,8]

Despite the benefits of using Rasch analysis to evaluate clinical scales, only one previous study has evaluated the nine-item ESAS[1] in a large sample ($n = 26{,}645$) of oncology outpatients.[9] Cheifetz et al.[9] confirmed the unidimensionality of the ESAS[1] and observed that patients reported their symptoms at the extreme low or high end of the scale. Further observations included that patients were not using responses 2–7.[2,9] This suggests that the ESAS[1] may have disordered item thresholds. This polarity in scoring concerns clinicians as it suggests that moderate symptoms are not being effectively assessed using the ESAS.[1]

The accurate assessment of symptom burden in palliative care is an essential component of clinical practice. This study aims to 1) evaluate the psychometric properties of the ESAS[1] using Rasch analysis in a sample of community hospice patients, 2) determine whether the performance of the ESAS[1] may be improved through scale modifications (e.g., item rescoring), and 3) to produce ordinal-to-interval Rasch conversation tables for clinical use that improve accuracy of measurement.

Methods

Patients and Settings

All patients enrolled in a community hospice service in New Zealand from December 2015 to May 2016 were included in this study. Two hundred twenty-nine ESAS[1] questionnaires from admission were included in the analysis. The ages of patients ranged from 25 to 99 years (mean 70.86 years; SD 14.04; 96 males and 133 females). Patients had a moderate level of symptom burden on the ESAS-Symptom Distress Score (SDS; mean 30.99; SD 15.00) and a moderate performance status on the Eastern Cooperative Oncology Group (mean 2.07; SD .92). Table 1 shows the participant demographics. About 95% of ESAS were completed by the registered nurse as a proxy for the patient.

Instruments

The ESAS[1] is a self-report numeric rating scale to assess symptom intensity on a scale between 0 (least degree of symptom) to 10 (worst degree of symptom). The ESAS[1] consists of 12 questions evaluating symptoms commonly associated with cancer—pain, tiredness, nausea, depression, anxiety, drowsiness, appetite, well-being, shortness of breath, constipation, insomnia, and complexity of care (staff completion), as well as an option of an open rater-selected symptom. Scoring can be assessed individually for each symptom or tallied

Table 1
Person Groups for DIF Analysis

Gender	Age Groups	Ethnicity Groups	Diagnosis Groups
Males ($n = 96$)	Younger than 65 yrs ($n = 75$)	Caucasian ($n = 154$)	Lung cancer ($n = 43$)
Females ($n = 133$)	66–80 yrs ($n = 88$)	Maori ($n = 22$)	Colorectal cancer ($n = 28$)
	81 yrs and older ($n = 66$)	Pacifica ($n = 30$)	Gastrointestinal cancer ($n = 24$)
		Asian ($n = 13$)	Gynecological cancer ($n = 16$)
		Other ($n = 10$)	Breast cancer ($n = 14$)

DIF = differential item functioning.
Other cancer ($n = 72$).
Noncancer diagnosis ($n = 32$).

to get a total SDS (ESAS-SDS). The ESAS can either be administered as a self-report or with the use of a proxy rater.

Procedure

Ethical approval was granted by the authors' Institutional Ethics Committee and the Internal Ethics Board at the community hospice where the research was undertaken. All data were deidentified on collection. Specific informed consent was not required as all participants gave signed consent for the use of their health information on admission to hospice; the ESAS[1] was part of their routine care.

Statistical Analysis

Rasch analysis was undertaken using RUMM2030 (RUMM Laboratory Pty, Ltd., Duncraig, WA) in the following sequential steps:[10]

1. An overall test of how well the data fitted the Rasch model.
2. Rescoring of items affected by disordered thresholds.
3. Reanalysis for overall model and individual item fit.
4. Examine local dependency and combine locally dependent items into superitems.
5. Further analysis for overall model and individual item fit.
6. Evaluation of DIF for age, sex, ethnicity, and diagnosis.
7. Test for unidimensionality.
8. Inspection of the residual correlation matrix for evidence of local dependency.
9. Examination of the person-item threshold distribution.

Rasch analysis is an iterative procedure whereby initial fit statistics are undertaken to ensure the overall fit of the data to the model. Any items that showed disordered thresholds were identified from the threshold map of RUMM2030 output. A disordered threshold reflects when individuals higher in the trait or ability being measured (symptom burden) do not consistently endorse higher response options for that item. Disordered thresholds may be corrected in Rasch analysis through collapsing adjacent response categories until ordered thresholds are found.[6] After each step, the fit to the model is retested.[7]

Item and person residuals provided summary fit statistics at a scale level. Perfect fit would have a mean of zero and an SD of one. A nonsignificant item-trait interaction Chi-squared fit statistic is desirable to demonstrate measurement consistency across different trait levels. Fit residuals at the individual item level should be around ±2.5, and Chi-squared statistics should be nonsignificant (>0.05 Bonferroni adjusted). Local dependency was evaluated through examining the residual correlation matrix for any high correlations among item residuals (>0.20) that might indicate local dependency.[11] Local response dependency occurs when responding on one item is influenced by response to another item.[12] Local response dependency can be remediated through the creation of superitems that combine scores of locally dependent items followed by retesting the model fit.[13] Evaluation of DIF was undertaken for age, ethnicity, and diagnosis (Table 1).

Smith[14] recommended evaluating the possibility of multidimensionality through inspecting the first principal component of the residuals after the Rasch factor has been removed. Person-locations are then compared using an independent t-test based on two subsets of items—one with the highest positive and the other with the highest negative loadings on the first principal component. Unidimensionality is established if less than 5% of t-test comparisons are significant. A binomial CI for proportions is also applied, and the lower bound must overlap 5%.

Results

Initial Test of the Model

The likelihood ratio test conducted with the sample of 229 patients (Table 2) in RUMM2030 software was significant ($P < 0.001$), which rejected the rating scale model and indicated the appropriateness of the partial credit Rasch model for the current analysis. The initial Rasch analysis showed a person separation index of 0.77, which indicated satisfactory reliability, but the overall model fit was poor ($\chi^2 = 155.41$ [108]; $P < 0.01$), and all items displayed disordered thresholds. The individual item fit

Table 2
Summary of Overall Rasch Model Fit Statistics

	Item Residual		Person Residual		Goodness of Fit		PSI	Independent t-Test	
Analysis	Mean	SD	Mean	SD	χ^2 (df)	*P*	Value	%	95% CI
1	0.34	1.44	−0.20	1.16	155.41 (108)	0.01	0.77	4.80	1.98
2	0.22	1.20	−0.32	1.21	142.28 (108)	0.02	0.75	6.55	3.73
3	0.22	1.19	−0.29	1.08	29.56 (27)	0.33	0.72	5.68	2.85

PSI = Person Separation Index; df = degrees of freedom.

Table 3
Individual Item Fit From the Initial Rasch Analysis of the 12-Item ESAS

Item	Item Difficulty (Location)	Item-Fit Residuals	χ^2
1. Pain	−0.13	2.76	12.55
2. Tiredness	−0.37	0.12	4.92
3. Nausea	0.29	0.21	6.84
4. Depression	0.59	0.65	9.63
5. Anxiety	0.12	0.16	8.04
6. Drowsiness	0.09	−1.44	8.43
7. Appetite	−0.17	0.46	4.63
8. Well-being	−0.37	−1.59	27.84
9. Shortness of breath	0.11	1.86	21.36
10. Complexity	−0.40	−1.21	20.21
11. Constipation	−0.01	2.51	21.34
12. Insomnia	0.25	−0.34	9.62

ESAS = Edmonton Symptom Assessment System.

was reviewed, and none of the items showed a misfit to the Rasch model (Table 3). Fig. 1 provides an example of a typical item category probability curve, displaying a disordered threshold.

Item Rescoring

All items showed disordered thresholds with similar patterns; therefore, optimal ordering of thresholds was achieved through uniform rescoring of all 11-point ESAS[1] items through collapsing response categories to a four-point scale (0 = none, 1 = mild, 2 = moderate, and 3 = severe). The first and the last response options were retained as they represented extreme conditions; for example, no pain and worst possible pain. Fig. 2 represents the item category probability curve for Item 1 (pain) after rescoring, illustrating perfectly ordered thresholds. Fig. 3 shows the threshold map for all ESAS[1] items after rescoring, ordered by degree of difficulty. There are no disordered thresholds, with complexity and well-being as the easiest to endorse items and drowsiness and depression as the most difficult to endorse items. The overall model fit was reassessed, but item-trait interaction Chi-square fit statistic was still significant ($\chi^2 = 142.2$ [108]; $P = 0.02$).

Examination for Local Dependency

The overall Rasch model fit can be affected by local dependency between items.[15] The residual correlation matrix indicated local dependency between six items forming three independent pairs. All residual correlations exceeded the 0.20 cut-off point above the mean of the residual correlations. The highest correlation was found between depression and anxiety (Items 5 and 6), followed by well-being and complexity (Items 8 and 10) and drowsiness and appetite (Items 6 and 7). Locally dependent items were combined into three subtests.[12] Good model fit was found on re-evaluation ($\chi^2 = 29.5$ [27]; $P = 0.33$), and strict unidimensionality was confirmed.

Differential Item Functioning

DIF was not found when evaluated by gender, age, ethnicity, and diagnosis.

Person-Item Threshold Distribution

The person-item threshold distribution of the rescored ESAS is shown in Fig. 4. Person ability (level of symptom burden) and item difficulty are plotted on the same interval logit scale. The plot shows that a broad range of symptom burden is assessed by the items of the ESAS, and there are no floor or ceiling effects. However, the patients who were assessed using the ESAS in this sample did not cover the range of abilities available from the ESAS; they were of a mild to moderate symptom severity. This is an advantage for a clinical scale as it enables the assessment of patients with a more extreme symptom burden than available in the present sample.

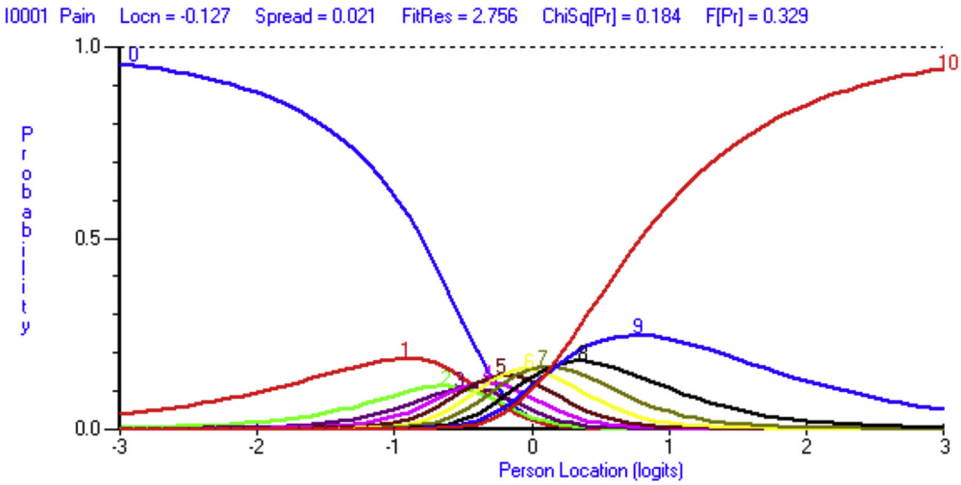

Fig. 1. Item category probability curves for Item 1 of the Edmonton Symptom Assessment System, before rescoring.

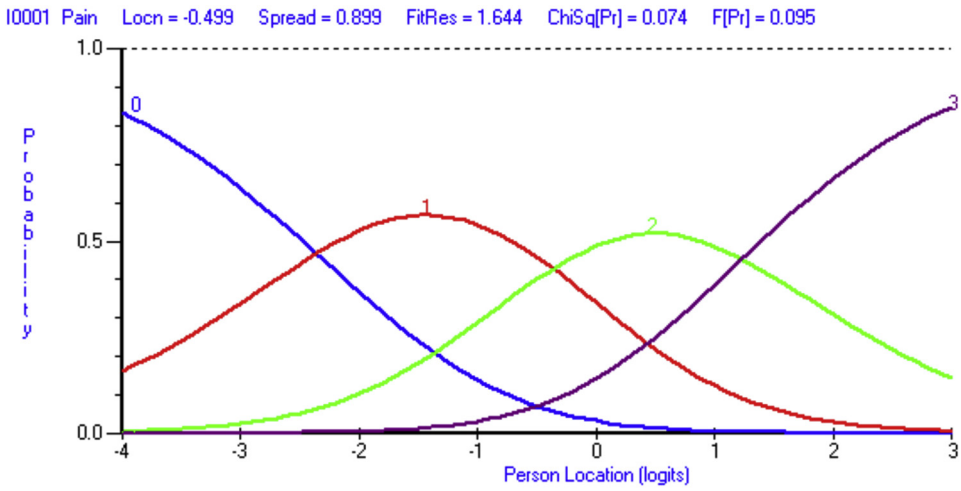

Fig. 2. Item category probability curves for Item 1 of the Edmonton Symptom Assessment System, after rescoring.

Ordinal-to-Interval Conversion

An ordinal-to-interval conversion table (Table 4) was produced to enable ordinal scores to be transformed to interval-level data without modifying the original response format of the scale. This table can only be used when the completed data for all items are available for the assessed individual.

Discussion

The ESAS[1] is a commonly used measure of symptom burden in palliative care and consequently requires robust psychometric properties. The results of the Rasch analysis of the ESAS[1] in a community palliative care setting support the use of the ESAS as a global unidimensional measure of symptom distress, albeit with some modifications to the scoring structure.

The satisfactory Rasch model fit and unidimensionality of the ESAS[1] found in this study is consistent with Cheifetz et al.[9] who also established that the nine-item version of the ESAS[1] fitted the requirements for the Rasch model and measured the same overall construct when used in a population of ambulatory cancer center patients. This is a significant finding as the present research took place in a population with a higher level of symptom burden than the study by Cheifetz et al.,[9] suggesting that the ESAS[1] retains good psychometric properties across a range of severity levels of patients. The person separation index of 0.72 (interpreted similar to Cronbach's alpha) indicates satisfactory reliability[6] of the ESAS in discriminating between the overall severity of symptoms.[1]

Rescoring

As found in the study by Cheifetz et al.,[9] this study found that responses to the ESAS[1] were clustered around the high and low ends of the scale, suggesting that raters were not discriminating between the 11 response options (0–10) of the ESAS.[1]

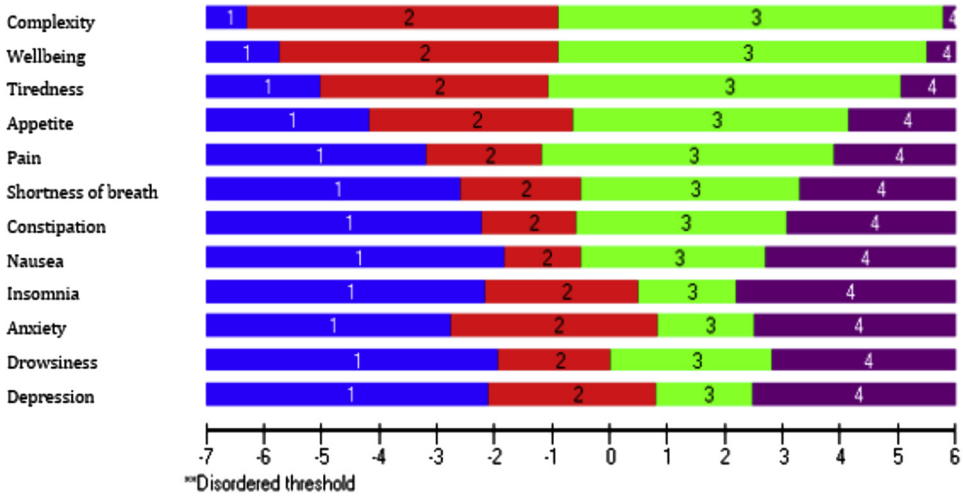

Fig. 3. Threshold map for the 12 Edmonton Symptom Assessment System items after uniform rescoring, ordered by difficulty.

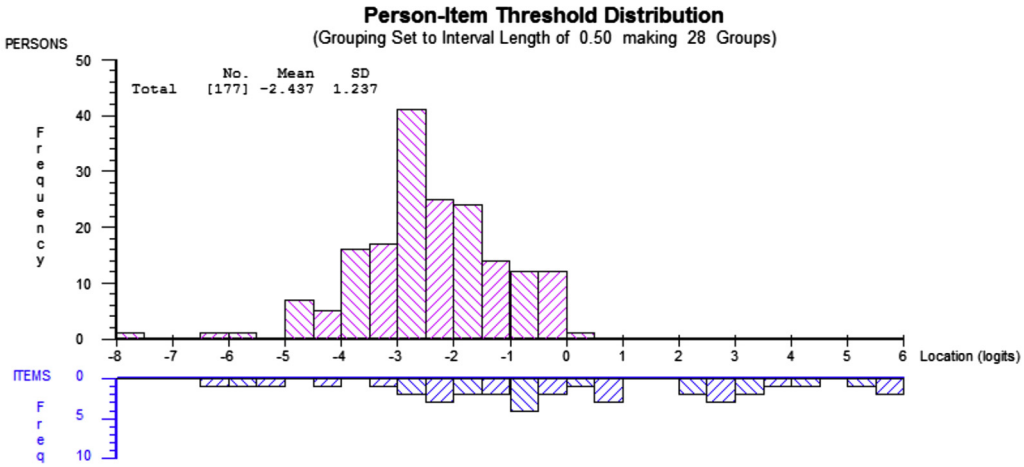

Fig. 4. Person-item threshold distribution.

Table 4
Conversion of Ordinal Scores Into Interval-Level Data, Logits, and Scale Metric for the Modified ESAS

Original ESAS Score After Rescoring (Ordinal Score)	Rasch Score Converted to Interval (Logits)	Converted ESAS-SDS Score Interval (Metric)
0	−8.64	0
1	−5.69	7
2	−4.21	11
3	−3.47	13
4	−2.92	14
5	−2.49	16
6	−2.13	16
7	−1.84	17
8	−1.58	18
9	−1.35	18
10	−1.13	19
11	−0.94	19
12	−0.75	20
13	−0.58	20
14	−0.42	21
15	−0.27	21
16	−0.14	21
17	−0.01	22
18	0.12	22
19	0.23	22
20	0.35	23
21	0.47	23
22	0.60	23
23	0.73	24
24	0.86	24
25	1.01	24
26	1.16	25
27	1.31	25
28	1.46	25
29	1.62	26
30	1.78	26
31	1.98	27
32	2.23	27
33	2.64	28
34	3.32	30
35	4.28	33
36	5.63	36

ESAS = Edmonton Symptom Assessment System; SDS = Symptom Distress Score.
Note: Rescoring of response options to individual items is required before converting ESAS-SDS scores into interval scale. See *Table 5* for values. Rescored sum of total scores should then be identified in the first column, and the final score should be located in the right column.

The present study undertook psychometric enhancement of item scoring of the ESAS[1] using Rasch analysis. Clinical use of the ESAS[1] using the collapsed four-point scoring (none, mild, moderate, and severe) responses suggested by this study may increase both reliability and clinical utility when compared with the previous 11-point scoring options. This extends the previous study undertaken by Cheifetz et al.[9] as our conversion table accounts for local dependency between items and uses the rescored scale metric. This facilitates both user friendliness and a higher precision of transformed scores. Clinicians should administer the ESAS using the 11-point scale; however, when data are analyzed for statistical purposes, rescoring to a four-point scale should occur. Alternatively, a global revision of the ESAS[1] could be undertaken to provide four item response options for the ESAS.

Local Dependency

Three pairs of items showed significant residual correlations indicative of local dependency. This suggests that the scores on these items influence each other and should be interpreted with caution. The highest residual correlation was observed between the depression and the anxiety items followed by the (absence of) well-being and complexity and drowsiness and the appetite item pairs. These pairs of symptoms may be expected in clinical practice. For example, moderate positive correlations between depression and anxiety have been consistently observed in th psychological literature.[16]

Differential Item Functioning

The evaluation of DIF revealed that there is no disadvantage in using the ESAS[1] across different population groups, regardless of if it is the nurse or client completing the ESAS.[1] This is aligned with the

Table 5
Rescored Item Response Values for the ESAS

Original Item Response Value	Rescored Item Value
0	0
1	1
2	
3	
4	
5	
6	2
7	
8	
9	
10	3

ESAS = Edmonton Symptom Assessment System.

findings of Cheifetz et al.,[9] who reported no DIF for age, sex, and diagnosis, in a group where patients were the predominate rater.

Person-Item Threshold Distribution

The ESAS items covered a range of symptoms from absent to severe, despite the population sampled mostly falling in the mild to moderate range of symptom burden (Fig. 4). This is advantageous for a clinical scale as patients who have higher symptom burden than the present sample, such as those who are deteriorating or at a later stage in their palliative care, will be able to be assessed using the ESAS.

Ordinal-to-Interval Conversion

Satisfactory fit to the Rasch model and unidimensionality of the ESAS[1] enables raw scores on the ESAS[1] to be transformed into interval scale estimates. Clinicians using the ESAS[1] can use the table provided to transform scores on the unmodified ESAS[1] to first, specify that a patient's condition has changed, and second, identify by how much, using a properly constructed interval scale (Table 4). These conversions improve accuracy of assessment, which is vital when data are being used to make clinical decisions.

Limitations

As data were collected during admission to a community hospice service, the range of symptoms exhibited may be at a lower severity when compared with a group of scores taken at regular intervals throughout their admission. Furthermore, the large proportion of ESAS[1] measures completed using proxy scoring by registered nurses in the present sample, the underrepresentation of younger aged patients, and lack of ethnic diversity restricted the ability to fully evaluate DIF in terms of patient responses. Consequently, future evaluation of DIF should be conducted on data where patients had completed the ESAS[1] with a broader age and ethnically diverse sample. Future research with a broader range of patient severity may investigate the performance of the ESAS[1] in patients at the extreme highs and lows of symptom burden and evaluate the performance of the ESAS[1] in different ethnic groups.

Conclusion

The present study has demonstrated that the ESAS[1] satisfies the expectations of the unidimensional Rasch model in a 12-item format after minor modifications. The psychometric properties and clinical utility of the ESAS[1] were significantly improved by uniform rescoring of disordered response categories and creating superitems. The accuracy of the ESAS[1] scores can be further improved by using ordinal-to-interval conversion tables published in the article.

Disclosures and Acknowledgments

The authors acknowledge the hospice staff and participants whose assessments provided data for this study. This research did not receive any specific grant from the funding agencies in the public, commercial, or not-for-profit sectors. The authors declare no conflicts of interest.

References

1. Bruera E, Kuehn N, Miller MJ, Selmser P, Macmillan K. The Edmonton Symptom Assessment System (ESAS): a simple method for the assessment of palliative care patients. J Palliat Care 1991;7:6–9.

2. Richardson LA, Jones GW. A review of the reliability and validity of the Edmonton Symptom Assessment System. Curr Oncol 2009;16:55.

3. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measure for clinical trial in neurology: problems, solutions, and recommendations. Lancet Neurol 2007;6:1094–1105.

4. Masters GA. Rasch model for partial credit scoring. Psychometrica 1982;47:149–174.

5. Curt H, Malin N, Gustavsson J. Using the Rasch model in nursing research: an introduction and illustrative example. Int J Nurs Stud 2009;46:380–393.

6. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007;57:1358–1362.

7. Medvedev O, Siegert R, Feng X, Billington D, Jang J, Krägeloh C. Measuring trait mindfulness: how to improve the precision of the mindful attention awareness scale using a Rasch model. Mindfulness 2016;7:384–395.

8. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol 1996;49:711–717.

9. Cheifetz O, Packham TL, Macdermid JC. Rasch analysis of the Edmonton Symptom Assessment System and research implications. Curr Oncol 2014;21:e186–e194.

10. Siegert R, Tennant A, Turner-Stokes L. Rasch analysis of the Beck Depression Inventory-II in a neurological rehabilitation sample. Disabil Rehabil 2010;32:8–17.

11. Marais I, Andrich D. Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. J Appl Meas 2008;9:105–124.

12. Lundgren-Nilsson Å, Grimby G, Ring H, et al. Cross-cultural validity of functional independence measure items in stroke: a study using Rasch analysis. J Rehabil Med 2005; 37:23–31.

13. Wainer H, Kiely G. Item clusters and computerized adaptive testing: a case for testlets. J Educ Meas 1987;24: 185–201.

14. Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3: 205–231.

15. Wright BD. Local dependency, correlations and principal components. Rasch Meas Trans 1996;10:509–511.

16. Clark L, Watson D. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. J Abnorm Psychol 1991;100:316–336.