Received: 19 July 2021 | Revised: 29 November 2022 | Accepted: 30 November 2022

DOI: 10.1002/cpp.2808

RESEARCH ARTICLE

WILEY

Examining the psychometric properties of the Swedish version of Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM) in a clinical sample using classical test theory and item response theory

Elisabeth Åström[1] | Anna E. Sundström[1] | Per-Erik Lyrén[2]

[1]Department of Psychology, Umeå University, Umeå, Sweden

[2]Department of Applied Educational Science, Umeå University, Umeå, Sweden

Correspondence
Elisabeth Åström, Department of Psychology, Umeå University, Umeå SE-901 87, Sweden.
Email: elisabeth.astrom@umu.se

Abstract

The aim of this study was to examine the psychometric properties of the Swedish version of the Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM) using classical test theory and item response theory (IRT). The CORE-OM is a commonly used 34-item self-report instrument measuring psychological problems/distress covering four domains: subjective well-being, problems/symptoms, functioning and risk. Despite its broad application, only a few studies have used IRT to examine the psychometric properties, and the properties of the Swedish version have only been examined in one initial study. The present study included 1,011 clients with mild to moderate symptoms of distress, applying for psychotherapy at an outpatient training clinic in Sweden. Clients' responses were subjected to classical item analyses as well as IRT (Rasch) analysis using the partial credit model. The classical analyses demonstrated high levels of internal consistency and acceptable levels of item discrimination for the majority of the items, although lower for some items, particularly in the Risk domain. IRT analyses showed that there was a rather good match between item and respondent locations and the measurement precision was high. Disordered step and average measures for some of the items in the Risk domain indicate that these items were problematic from a psychometric point of view and only applicable for a minority of the participants. Differential item functioning for gender in some of the items suggests that they might need to be revised to minimise potential gender bias.

KEYWORDS
classical test theory, CORE-OM, item response theory, Rasch analysis

1 | INTRODUCTION

Self-report instruments are often used to assess psychological distress and psychopathology in research as well as clinical practice. To draw accurate conclusions, a basic requirement is that the instruments used for this purpose are reliable and valid; i.e., scores should accurately reflect clients' standing on the measured dimensions. The Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM;

Evans et al., 2000) is frequently used in both research and in clinical practice to measure psychological distress (Evans et al., 2000) and to evaluate the effects of psychological treatment (Evans et al., 2017). CORE-OM includes 34 items covering four domains: Subjective well-being, measuring general well-being; Problems/symptoms, which measures symptoms of anxiety, depression, trauma, and somatic complaints; Functioning, which comprehends interpersonal, social and general functioning; and, finally, Risk, which covers both risk to self (such as, suicidal ideation and self-harm) and risk to others (threats and physical violence). The items are described as being of both low and high intensity, meaning that they are constructed to capture a wide range of distress (Barkham et al., 2006). Items belonging to the Risk domain generally have the highest intensity; i.e., they can capture the most severe cases of distress or psychopathology.

CORE-OM was developed to be a broad and pan-theoretical measure, to facilitate use by clinicians working from different theoretical perspectives and in different settings and to be comprehensible enough to be implemented as a tool for routine evaluation in mental health and counselling services (Barkham et al., 2006; Evans et al., 2000). The items in the CORE-OM were specifically chosen to be of high clinical utility and to be acceptable to patients (Evans et al., 2000). Due to its broad approach, CORE-OM is not restricted to specific diagnostic groups or symptom clusters, in contrast to more symptom specific measures. Although there are other, similarly broad measures of global distress, such as the Symptom Checklist-90 (Derogatis et al., 1973) or its short form, the Brief Symptom Inventory (Derogatis & Melisaratos, 1983), CORE-OM holds an advantage in that it is relatively short. It is also free to use without a fee, provided that it is not altered in any way (see coresystemtrust.org.uk/home/copyright-licensing/). This has probably facilitated its widespread use; to date CORE-OM has been translated to over 30 languages, including Swedish. The widespread use of CORE-OM can further enable international comparisons and multicenter studies (Zeldovich et al., 2019).

1.1 | Psychometric properties of CORE-OM

Psychometric properties of self-report instruments are often investigated using classical test theory methods (CTT; Crocker & Algina, 1986). Item response theory (IRT) methods can contribute with important information in examining the quality of items and the functioning of the scale (Hambleton et al., 2000). Although several studies have examined the psychometric properties of CORE-OM using CTT, there are only a few studies, to our knowledge, that have applied IRT to evaluate CORE-OM (e.g., Bedford et al., 2010; Murray et al., 2014; Zeldovich et al., 2019).

IRT-based models are a collection of statistical models that describe the relationship between a latent construct/trait (i.e., what the scale is constructed to measure), the characteristics of the items in the scale and individuals' responses to the scale's items (Embretson & Reise, 2000). In essence, responses to items are a function of the properties of the items and the person's level on the latent construct. IRT has been used extensively for psychometric evaluation of tests in the educational field. However, several scholars have recognised the value of IRT in clinical assessment and patient reported outcome measures (e.g., Nguyen et al., 2014; Thomas, 2011), and the use of IRT within clinical research has increased in recent years (Thomas, 2019). There are several benefits of using IRT to evaluate rating scales, provided that the IRT model used fits the data (see e.g., van der Linden & Hambleton, 1997). First, item statistics are independent of the persons examined, which implies that the psychometric properties of an instrument are not sample-dependent and scores describing respondents' standing on the underlying trait measured are not test-dependent. Second, item location and person trait level are indexed on the same metric, which means that a scale can be effectively targeted to the population that the scale was constructed for. Third, item and test/scale information are used to assess the precision of measurement provided at points along the latent construct continuum. Consequently, IRT can be used for developing effective rating scales since the efficiency of the scale can be adapted to a specific level of the construct (Hambleton et al., 2000).

For the IRT-modelling in the current study, we used a Rasch model for polytomous data. In the Rasch model, the relationship between the latent trait and the items is indicated by the difficulty or threshold parameter (referred to as beta; *b*). This parameter indicates the probability of a person with a certain level of the latent trait (theta; θ) to endorse an item with difficulty or threshold *b*. To perform a Rasch analysis, three basic assumptions (Kean et al., 2018) need to be met: (1) unidimensionality (all test items measure the same latent construct); (2) local item independence (responses to items are not dependent on one another, after controlling for the latent trait); and (3) monotonicity (the probability of higher scores on an item is corresponding to higher levels on the latent trait). A careful analysis of the assumptions allows the researcher to evaluate how well data fits the model, that is, the measurement validity (Hambleton et al., 1991; Wilson, 2005). An advantage of Rasch methodology is that item bias between different subpopulations (e.g., men and women or individuals belonging to different ethnic groups) is relatively easily examined. In the Rasch model, item bias can be assessed using the differential item functioning analysis (DIF). DIF will occur when individuals from

Key Practitioner Messages

- Results from classical and IRT analysis indicated that the CORE-OM is a measure with high reliability for measuring psychological distress.
- Analysis of the functioning of the rating scale showed that items in the Risk domain might be relevant for only a minority of patients.
- Some of the items displayed differential item functioning for gender and may not be equally endorsed by men compared to women despite the same levels of distress.

different subpopulations, but with the same latent trait level, do not have the same probability of endorsing a specific item.

1.2 | Prior psychometric evaluations of CORE-OM using classical test theory

The original as well as translated versions of CORE-OM have been subject to a number of psychometric studies examining aspects of its reliability and validity in different settings and samples (see, e.g., Evans et al., 2002; Zeldovich & Alexandrowicz, 2019). The results of these studies indicated high levels of internal consistency as well as stability (Evans et al., 2002). Large score differences between clinical and non-clinical samples suggest that CORE-OM can discriminate between clinical and non-clinical groups (Elfström et al., 2013; Evans et al., 2002). The instrument has further demonstrated good convergent validity with instruments tapping closely related variables, such as depression, anxiety and other symptoms of psychological ill-health (Elfström et al., 2013; Evans et al., 2002; Palmieri et al., 2009).

The construct validity and dimensionality of CORE-OM have been examined in several factor analytic studies with somewhat mixed results (e.g., Evans et al., 2002; Lyne et al., 2006; Skre et al., 2013). As the main focus of the present study was to examine the functioning of the CORE-OM using a Rasch model, it was of specific interest whether CORE-OM was sufficiently unidimensional (one-factorial) to be subjected to a Rasch analysis. It was clear from a recent systematic review and meta-analysis examining the psychometric properties (including the factorial structure) of CORE-OM (Zeldovich & Alexandrowicz, 2019) that the four domains of CORE-OM do not map onto four separate factors. Instead, studies have reported one, two or three-factorial solutions (for a review, see Zeldovich & Alexandrowicz, 2019).

In the development of CORE-OM, the ambition was to minimise gender bias (Evans et al., 2000). Subsequent research has shown that there are gender differences in the CORE-OM domains, although they differ in magnitude between clinical and non-clinical groups and between different translations and countries. In the UK (and original) version, there were moderate and significant differences in all domains except Functioning non-clinical samples, whereas differences in clinical samples were only significant for subjective well-being and risk (Evans et al., 2002). The Swedish version (Elfström et al., 2013) showed significant gender differences on problems/symptoms in a non-clinical sample, but significant differences on both subjective well-being and risk in a clinical sample. In an Italian, clinical sample, more pronounced gender differences were found than in the UK sample, apart from on the Risk domain (Palmieri et al., 2009). In sum, previous studies indicate that there might exist gender differences in the CORE-OM items, although they may differ between the domains, samples (clinical vs. nonclinical) and countries. From a validity standpoint, however, it is important to examine whether such differences are due to item bias or if they reflect real differences between men and women.

1.3 | Prior psychometric evaluations of CORE-OM using IRT

There are only a few prior psychometric evaluations of CORE-OM using IRT-methods. In the first study (Bedford et al., 2010), Mokken scaling, a nonparametric IRT technique that has been used for analysing invariant item ordering in clinical scales (see, e.g., Watson et al., 2012), was utilised. Two separate analyses were made: one including the full inventory and another analysis excluding the Risk domain. In the analysis of the full scale, six items had acceptable scaling properties in measuring the latent trait on a reasonable hierarchy of distress and risk. In this model, items ranged in severity of distress/risk from *I have achieved the things I wanted to do* (easiest item, requiring the least level of the latent trait) to *I made plans to end my life* (most difficult item, requiring the highest level of the latent trait). In the analysis excluding the risk items, eight items had acceptable scaling properties and ranged in severity of distress from *I have felt tense, anxious or nervous* (easiest item) to *I have felt panic or terror* (most difficult item).

In the second study (Murray et al., 2014), Mokken scaling was used to investigate invariant item ordering of the CORE-OM before and after treatment (i.e., whether items and item response categories measured the same level of the latent trait). In short, 10 of the items demonstrated invariant item ordering before and after treatment (Murray et al., 2014), which implied measurement stability and the property of these items to reliably measure change over the course of treatment.

In the third study (Zeldovich et al., 2019), multidimensional IRT-modelling was applied to examine several aspects of CORE-OM, including the functioning of the response scale and DIF (with respect to gender and diagnostic groups). The authors found that the five-category response scale might be over-differentiating, in that the extreme response category (most or all of the time) was rarely used in a majority of the items. In the DIF analysis, at least five items exhibited DIF with respect to gender. Despite the same latent trait level, women were more likely to endorse three of those items (items 2, 14 and 20), whereas two items were more likely to be endorsed by men (items 16 and 19). This led the authors (Zeldovich et al., 2019) to conclude that separate norms might be needed for women vs. men.

1.4 | The Swedish version of CORE-OM

The psychometric properties of the Swedish version of CORE-OM was examined in only one prior study (e.g., Elfström et al., 2013). The Swedish version demonstrated acceptable to high levels of internal consistency and stability across the full scale and within the domains (Elfström et al., 2013). Convergent validity was examined by correlating CORE-OM scores with the Hospital Anxiety Depression Scale (Zigmond & Snaith, 1983) and the results showed correlations in expected directions. The factor structure of the Swedish instrument was not examined in the study (Elfström et al., 2013); however,

moderate to high intercorrelations between the domains indicate that they were not completely independent.

1.5 | Aim of the current study

In the current study, we evaluated the psychometric properties of the Swedish CORE-OM. Reliability was examined using CTT methods, but the central aim was to evaluate the psychometric properties of CORE-OM using IRT (Rasch) analyses. In this regard, we examined item and person location to obtain information of whether CORE-OM was a suitable instrument targeted to the client group. The ordering of response categories was assessed to see whether the rating scale functioned properly. Finally, the items were examined for DIF with respect to gender. The study took place at an outpatient training clinic and included a large sample of clients ($N = 1{,}011$) with mild to moderate symptoms of distress.

2 | METHOD

2.1 | Participants

The sample included 1,011 clients who had applied for psychotherapy at the Psychology Clinic, an outpatient training clinic at a university in Sweden. Only pre-treatment data were used in the current study. The mean age was 28 years ($SD = 8.8$; range: 19–91 years), and 719 (71%) were women. The clients had mild to moderate symptoms of distress. As the psychology clinic is a training clinic for students at the Clinical Psychology programme, persons with more severe psychopathology, including severe eating disorders, severe depressive symptomatology, suicidality, chronic problems, severe co-morbidity and low functioning level are not admitted to the clinic. The clients' primary complaints were identity/self-image problems (78%), relationship problems (71%), anxiety (69%), depression (61%), crisis/stress (33%), phobia (20%) and other difficulties. Altogether, 90% reported psychological problems in more than two areas.

2.2 | Procedure

Clients were self-referred to the clinic through an Internet website and experienced clinical psychologists at the clinic performed intake interviews that included a clinical assessment. During the interviews, clients were asked to participate in the study, and written informed consent was obtained from those who chose to do so. Before treatment was initiated, the clients filled out a computer-based questionnaire, consisting of several questionnaires including the CORE-OM. The study followed the Declaration of Helsinki and was approved by the Regional Ethics Board at the authors' institution. The data used in this study came from a project entitled 'Outcome and Prediction of Outcome in Psychotherapy Training Programs', and data were collected between 2012 and 2015.

2.3 | Instruments

2.3.1 | Clinical outcomes in routine evaluation – Outcome measure

CORE-OM (Evans et al., 2000) is a self-report measure consisting of 34 items covering four domains: subjective well-being (4 items), problems/symptoms (12 items), functioning (12 items) and risk to self and others (6 items). The respondent is asked to answer the items according to how often they have felt that way during the last week. Items are scored on a 5-point Likert scale (0 = *not at all*; 1 = *only occasionally*; 2 = *sometimes*; 3 = *often*; 4 = *most or all the time*), where higher scores reflect a higher frequency of problems. Eight items have reversed scoring. The Swedish version of CORE-OM (Elfström et al., 2013) was used in the current study. Test–retest reliability for the full scale was .85 for the Swedish version (Elfström et al., 2013) and .90 for the original version (Evans et al., 2000). Internal consistency was $\alpha = .93$ for the Swedish version (Elfström et al., 2013) and $\alpha = .94$ for the original version (Evans et al., 2002).

2.4 | Statistical analyses

Classical item statistics including item mean, standard deviation, corrected item-total correlation and Cronbach's alpha were examined first. Dimensionality was then assessed using a Principal Component Analysis (PCA). The reason for including a PCA was to assess whether CORE-OM fulfilled the assumption of unidimensionality, which is a prerequisite for performing a Rasch analysis. A rule of thumb for evaluating whether the assumption of unidimensionality is met is that the first factor should be dominant and account for more than 20% of the variability, and the first eigenvalue should be at least four times the second eigenvalue (Hambleton, 2005). The assumptions of Local item independence and monotonicity were examined as part of the Rasch analyses, which were executed using the software Acer ConQuest (Wu et al., 1997). Local item independence was checked by assessing item fit between each item pair in the CORE-OM using ConQuest's FIT procedure. Local dependence was deemed present if item fit between pairs of items exceeded 1.33. Monotonicity was checked by visual inspection of item observed score curves depicting the respondents' item scores relative to their standing on the latent trait (by binning of the respondents). If respondents located higher on the latent trait (i.e., respondents with more severe symptoms) have a lower observed item score than respondents located at lower levels on the latent trait, it indicates that the monotonicity assumption could be violated.

Two commonly used one-parametric Rasch models for polytomous data, the partial credit model (PCM) and the rating scale model (RSM), were fitted to the data, and the fit of these models was compared. The PCM estimates the probability of answering in a specific response category and estimates a parameter that predicts the amount of trait needed to move from one response category to the

next. In the RSM, the amount of trait needed to move from one response category to the next is fixed to be the same for all items in the scale. The PCM gives more flexibility in fitting the data, whereas RSM is more parsimonious because it requires less parameters to be estimated (Bond & Fox, 2004).

Item parameters (thresholds) were estimated using marginal maximum likelihood estimation (ConQuest default setting). Model fit was examined using weighted fit (infit) and unweighted fit (outfit) values within the range of 0.75–1.33 as a criterion of a good fit (Wilson, 2005).

The functioning of the rating scale was evaluated according to established guidelines (e.g., Linacre, 1999, 2003). Average measures for the rating scale categories were examined because ordered average measures imply that higher categories on the rating scale correspond to higher levels of the latent construct (Linacre, 1999). Moreover, the ordering of step measures (delta values in ConQuest) was examined because ordered step measures indicate that as one moves up the latent trait continuum, each category in turn becomes the most probable response. Step measure parameters are at the intersections of adjacent category curves and are expected to monotonically advance with increasing participant ability (or, as is the case with clinical measures, symptom severity) (Linacre, 2003).

The test information function (TIF) was used to assess if the information across the test items matched the latent distribution (i.e., the location of the respondents' scores). TIF is the sum of the information functions for all items. Item information curves display the information as a function of trait level. An item provides more information when the item location parameter completely matches a person's trait location (Embretson & Reise, 2000). Test information is related to the standard error of the measure in the sense that the standard error is the inverse square-root of the test information. Thus, the more information provided by an instrument at a particular location, the smaller the errors associated with estimation of item and respondent location (Hambleton et al., 1991; Wilson, 2005). A common guideline is that test information around 10 provides adequate measurement precision, as it corresponds to a standard error of about 0.31, equivalent to a reliability of 0.90 (Embretson & Reise, 2000).

Finally, differential item functioning (DIF) analyses were used to examine whether men and women with similar levels of the latent trait have different probabilities of endorsing an item. We examined DIF both on an item level and step level (i.e., DIF for each response category, also referred to as differential step functioning). The items and response categories within each item were calibrated using gender as a covariate in the model and the item and step locations for men and women were compared. A value below .426 is considered negligible, values between .426 and .638 are considered intermediate DIF, and values over .638 are considered large DIF in items (Wilson, 2005). Similar thresholds can be used to examine DIF on a step level (Penfield & Gattamorta, 2009).

3 | RESULTS

3.1 | Classical test theory

Classical item statistics, including endorsement proportions, are reported in Table 1.

Cronbach's alpha was .92 for the total scale. There was no item that would meaningfully improve the reliability of the scale if deleted (see Table 1). Corrected item total correlations showed that items 6, 8, 22 and 34 had correlations below .30, indicating that these items were relatively weakly associated to the total score (see Table 1). Items 6, 22 and 34 belong to the Risk domain, and item 8 asks about pain, aches and other physical symptoms (Problems/Symptoms). All items had responses in all categories except item 6, although items 16, 22, 24, 33 and 34 had categories with less than 10 responses. Further inspection of the data showed that six of the items (all belonging to the Risk domain) were positively skewed (skewness > 1.5), indicating that very few participants had used the higher response categories.

3.2 | Item response theory

Before conducting the Rasch analysis, the dimensionality of the data was examined through a Principal Component Analysis (PCA). All 34 items in the CORE-OM were subjected to the PCA. The examination of eigenvalues and the scree plot indicated that one dominant factor was present. The first factor had an eigenvalue of 10.26, explaining 30% of the variation. Factor 2 had an eigenvalue of 2.12 and explained 6% of the variation. Thereby, it was concluded that the assumption of unidimensionality was fulfilled (Hambleton, 2005), and a Rasch model could be applied to the data. Because item 6 (*I have been physically violent to others*) did not have responses in all categories (see Table 1), this item was removed from further analyses.

The analysis of Local item independence showed that item 8 was the most problematic item and had dependency with 14 other items. Apart from item 8, some items belonging to the Functioning domain (items 3, 19, 21, 25, 26, 32 and 33) demonstrated dependency with items within that domain (see Table S1). Finally, the visual inspection of item observed score curves showed that there were no cases of severe violations of the monotonicity assumption, for example, curves that were monotonically decreasing, bell-shaped or uniform.

3.2.1 | Item fit, item location and rating scale functioning

Item fit was compared for RSM and PCM. For RSM, 10 items had unacceptable infit mean squares, whereas for PCM three items had mean-square infit values outside the acceptable boundaries. Item 8 (*troubled by aches, pains or other physical problems*) had an infit value of 1.63; item 19 (*felt warmth or affection for someone*) an infit value of

TABLE 1 Mean (M), standard deviations (SD), Cronbach's alphas (α), item-total correlation, skewness, kurtosis and endorsement proportions

							Endorsement proportions				
Item	M	SD	α if item omitted	Item-total correlation	Skewness	Kurtosis	Not at all	Only occasionally	Sometimes	Often	Most or all the time
1	1.70	1.12	.92	.66	.07	−.79	0.17	0.25	0.33	0.20	0.05
2	2.37	.99	.92	.55	−.31	−.37	0.04	0.15	0.33	0.37	0.11
3	1.38	1.12	.92	.37	.38	−.78	0.27	0.30	0.25	0.15	0.03
4	2.18	.94	.92	.60	−.17	−.32	0.04	0.19	0.40	0.32	0.07
5	2.06	1.07	.92	.62	.03	−.69	0.06	0.26	0.33	0.26	0.10
6	.03	.23	.92	.11	9.19	115.84	0.97	0.03	0.00	-	0.00
7	1.61	.96	.92	.47	.26	−.43	0.11	0.37	0.33	0.16	0.03
8	1.66	1.34	.93	.25	.26	−1.13	0.27	0.21	0.23	0.18	0.11
9	.29	.66	.92	.46	2.53	6.43	0.81	0.11	0.06	0.01	0.00
10	1.46	1.10	.92	.56	.34	−.61	0.23	0.30	0.30	0.13	0.04
11	1.57	1.19	.92	.58	.19	−.99	0.24	0.25	0.26	0.20	0.05
12	1.85	.92	.92	.58	.15	−.29	0.06	0.31	0.41	0.19	0.04
13	2.51	1.07	.92	.58	−.51	−.37	0.05	0.14	0.25	0.40	0.17
14	1.96	1.13	.92	.55	−.15	−.76	0.13	0.21	0.32	0.27	0.07
15	1.01	1.10	.92	.55	.89	−.10	0.43	0.28	0.18	0.09	0.03
16	.13	.44	.92	.38	4.25	21.22	0.90	0.07	0.02	0.01	0.00
17	1.38	1.22	.92	.72	.50	−.77	0.30	0.28	0.22	0.15	0.06
18	1.81	1.36	.92	.45	.19	−1.15	0.22	0.23	0.24	0.17	0.15
19	1.11	1.03	.92	.25	.78	.05	0.32	0.37	0.20	0.08	0.03
20	2.22	1.12	.92	.63	−.10	−.73	0.07	0.20	0.33	0.26	0.15
21	1.15	.93	.92	.45	.67	.04	0.25	0.46	0.19	0.09	0.01
22	.11	.43	.92	.15	5.06	31.09	0.92	0.06	0.01	0.00	0.00
23	1.82	1.18	.92	.75	.10	−.84	0.16	0.25	0.30	0.21	0.09
24	.42	.82	.92	.52	2.16	4.32	0.74	0.15	0.07	0.03	0.01
25	1.43	1.10	.92	.47	.38	−.62	0.24	0.31	0.28	0.14	0.04
26	1.13	1.19	.92	.43	.79	−.38	0.41	0.25	0.19	0.10	0.05
27	1.85	1.17	.92	.76	.03	−.85	0.15	0.24	0.31	0.23	0.08
28	1.82	1.25	.92	.53	.01	−1.04	0.20	0.20	0.28	0.23	0.09
29	1.43	1.09	.92	.43	.34	−.64	0.24	0.30	0.30	0.13	0.04
30	2.12	1.24	.92	.46	−.18	−.93	0.13	0.18	0.28	0.27	0.15
31	1.98	1.01	.92	.51	−.03	−.53	0.07	0.25	0.37	0.25	0.06
31	1.67	.89	.92	.54	.30	−.05	0.07	0.37	0.40	0.13	0.03
33	.64	.88	.92	.38	1.42	1.56	0.57	0.28	0.10	0.04	0.01
34	.12	.45	.92	.29	4.67	25.17	0.92	0.06	0.02	0.01	0.00

Note: $N = 1{,}011$; items are divided on four domains: Subjective well-being: 4, 14, 17, 31; Problems/symptoms: 2, 5, 8, 11, 13, 15, 18, 20, 23, 27, 28, 30; Functioning: 1, 3, 7, 10, 12, 19, 21, 25, 26, 29, 32, 33; Risk: 6, 9, 16, 22, 24, 34.

1.37; and item 27 (*felt unhappy*) an infit value of 0.69. Item 17 (*felt overwhelmed by my problems*) was borderline misfit at 0.74. Based on the examination of item fit, the PCM was chosen for parameter estimation (see Table 2).

Item locations are displayed in Table 2. Based on the values for item location, item 13 (*been disturbed by unwanted thoughts and feelings*) was on average the easiest item for the participants to endorse, followed by item 2 (*felt tense, anxious or nervous*), whereas item 16 (*made plans to end my life*) was on average the most difficult item for participants to endorse followed by the rest of the items belonging to the Risk domain.

The functioning of the rating scale was evaluated by examining the ordering of step measures and average measures for the rating scale categories. The step measures for each item increased monotonically across rating scale categories for all items except for item 16 and 22 (see Table 2). With respect to average measures, most

TABLE 2 Rasch analyses: Average item location, step measures and item fit

Item	Average item location	Standard error	Step measures				Infit	Outfit
			T1	T2	T3	T4		
13. I have been disturbed by unwanted thoughts and feelings.	−1.16	0.05	−2.50	−1.63	−1.09	0.56	0.94	0.94
2. I have felt tense, anxious or nervous.	−1.07	0.05	−2.86	−1.74	−0.67	0.97	0.94	0.95
20. My problems have been impossible to put to one side.	−0.92	0.04	−2.41	−1.38	−0.28	0.40	0.89	0.88
4. I have felt ok about myself.	−0.86	0.06	−2.98	−1.64	−0.26	1.44	0.86	0.87
5. I have felt totally lacking in energy and enthusiasm.	−0.77	0.05	−2.67	−1.08	−0.19	0.88	0.88	0.88
30. I have thought I am to blame for my problems and difficulties.	−0.70	0.04	−1.54	−1.26	−0.46	0.46	1.17	1.20
31. I have felt optimistic about my future.	−0.58	0.05	−2.54	−1.18	−0.04	1.43	1.00	1.01
12. I have been happy with the things I have done.	−0.49	0.06	−2.94	−1.05	0.40	1.66	0.89	0.89
14. I have felt like crying.	−0.49	0.04	−1.69	−1.25	−0.24	1.22	0.99	1.00
18. I have had difficulty getting to sleep or staying asleep.	−0.47	0.03	−1.10	−0.76	−0.04	0.01	1.26	1.35
23. I have felt despairing or hopeless.	−0.42	0.04	−1.59	−0.93	0.01	0.82	0.70	0.70
27. I have felt unhappy.	−0.42	0.04	−1.58	−0.99	−0.09	1.00	0.69	0.69
28. Unwanted images or memories have been distressing me.	−0.38	0.04	−1.10	−1.06	−0.21	0.86	1.07	1.09
8. I have been troubled by aches, pains or other physical problems.	−0.30	0.04	−0.79	−0.71	−0.11	0.42	1.61	1.83
32. I have achieved the things I wanted to.	−0.27	0.06	−2.81	−0.76	0.81	1.68	0.93	0.92
1. I have felt terribly alone and isolated.	−0.21	0.05	−1.44	−0.97	0.18	1.40	0.83	0.83
7. I have felt able to cope when things go wrong	−0.12	0.06	−2.30	−0.54	0.46	1.89	1.04	1.06
11. Tension and anxiety have prevented me doing important things.	−0.08	0.04	−1.07	−0.69	−0.02	1.45	0.96	0.96
10. Talking to people has felt too much for me.	−0.02	0.05	−1.28	−0.60	0.57	1.24	0.96	0.95
17. I have felt overwhelmed by my problems.	−0.01	0.04	−0.84	−0.32	0.18	0.97	0.74	0.75
25. I have felt criticized by other people.	0.02	0.05	−1.26	−0.48	0.51	1.32	1.10	1.11
29. I have been irritable when with other people.	0.05	0.05	−1.20	−0.58	0.59	1.41	1.15	1.16
3. I have felt I have someone to turn to for support when needed.	0.12	0.05	−1.08	−0.38	0.29	1.67	1.26	1.29
26. I have thought I have no friends.	0.21	0.04	−0.37	−0.18	0.49	0.89	1.19	1.25
19. I have felt warmth or affection for someone.	0.32	0.05	−1.01	0.17	0.80	1.30	1.37	1.49
15. I have felt panic or terror.	0.40	0.04	−0.36	0.03	0.61	1.31	0.95	0.94
21. I have been able to do most things I needed to.	0.46	0.08	−1.50	0.39	0.71	2.22	1.04	1.06
33. I have felt humiliated or shamed by other people.	0.92	0.09	0.07	0.73	0.97	1.90	1.11	1.09
24. I have thought it would be better if I were dead.	1.04	0.08	1.05	0.58	0.93	1.61	0.86	0.69
9. I have thought of hurting myself.	1.34	0.12	1.44	0.49	1.89	1.54	0.88	0.71
22. I have threatened or intimidated another person.	1.52	0.13	2.31	1.45	1.61	0.72	1.12	1.68
34. I have hurt myself physically or taken dangerous risks with my health.	1.57	0.15	2.33	1.22	1.23	1.49	0.98	0.81
16. I made plans to end my life.	1.77	0.21	2.06	1.60	1.11	2.31	0.91	0.60

items average measure increased monotonically with rating scale category. For some of the items in the risk domain (items 9, 16, 24 and 34), however, average measures were unordered for some of the categories (item 6 removed from analyses). In common for all items with disordered step measures or average measures was that the distribution of scores was very restricted as only very few participants had responded in the three highest categories (see Table 1).

3.2.2 | Differential item functioning and test information

Differential item functioning (DIF) was examined with respect to gender. Inspection of responses of each response category (0 to 4) for women versus men, showed that men had no responses in category 4 for item 9 and no responses in category 3 for item 22, whereas women had no responses in category 4 for item 16. Therefore, items 9, 16 and 22 were removed from the DIF-analyses. Model fit was compared for the analyses where DIF was modelled on an item level versus on a step level. Model fit was better for the latter model and results from this analysis is thus reported in Table 3. Six of the items displayed DIF but none of the items displayed DIF across the whole response scale; instead, the potential bias was located in individual score categories (see bold markings in Table 3).

Lastly, we examined test information. The test information function peaked at 29 at 0 logits and the test information was around 15 between −2.0 and 2.0 logits (see Figure S1). The test information was less than 10 (corresponding to a reliability below α =. 90) above and below ±2.5 logits. The majority of the participants were located well within this interval, meaning that the test could accurately measure their level on the latent trait. However, and as can be seen in Figure 1, some of the items were located outside and at a higher trait level than the persons in the sample, illustrating that these items were too difficult for participants to endorse. On the contrary, Figure 1 also shows that some participants were located outside and at a lower trait level than the instrument captured.

4 | DISCUSSION

The aim of this study was to examine the psychometric properties of the Swedish translation of CORE-OM using CTT as well as IRT based methods. Beginning with the CTT analyses, the results showed that CORE-OM had excellent internal consistency (α = .92), similar to that reported by other studies (e.g., Holmqvist et al., 2014; Skre et al., 2013) and in line the previous study of internal consistency of the Swedish version of CORE-OM (Elfström et al., 2013). The item response distribution was also in line with the previous evaluation of the Swedish CORE-OM and showed that the Risk items were all positively skewed. The item-total correlation further showed that most items contributed meaningfully to the overall score (r_s > .30) and discriminated between participants with higher and lower levels of distress. However, item 6 (*been physical violent to others*), item 8 (*troubled by aches or pains or other physical problems*), item 19 (*felt warmth or affection for someone*), item 22 (*threatened or intimidated another*

TABLE 3 Items displaying differential item functioning

Item	Step	Women step location	Men step location	Difference step location
5. I have felt totally lacking in energy and enthusiasm.	Step 1	−1.802	−1.974	0.172
	Step 2	−0.135	−0.374	0.239
	Step 3	**0.274**	**0.713**	**0.439**
	Step 4	1.663	1.635	0.028
7. I have felt able to cope when things go wrong	Step 1	−1.980	−2.265	0.285
	Step 2	−0.187	−0.502	0.315
	Step 3	0.567	0.598	0.031
	Step 4	**1.599**	**2.168**	**0.578**
10. Talking to people has felt too much for me.	Step 1	−1.302	−1.249	0.053
	Step 2	−0.338	−0.681	0.343
	Step 3	**0.198**	**0.762**	**0.564**
	Step 4	1.442	1.169	0.271
11. Tension and anxiety have prevented me doing important things.	Step 1	−0.736	−1.092	0.356
	Step 2	**−0.939**	**−0.478**	**0.461**
	Step 3	0.193	−0.001	0.194
	Step 4	1.482	1.571	0.089
12. I have been happy with the things I have done.	Step 1	−2.327	−2.528	0.201
	Step 2	−0.751	−0.491	0.260
	Step 3	0.628	1.007	0.379
	Step 4	**2.451**	**2.012**	**0.439**
14. I have felt like crying.	**Step 1**	**−1.091**	**−1.546**	**0.455**
	Step 2	−0.780	−0.719	0.061
	Step 3	0.229	0.360	0.131
	Step 4	1.642	1.905	0.263

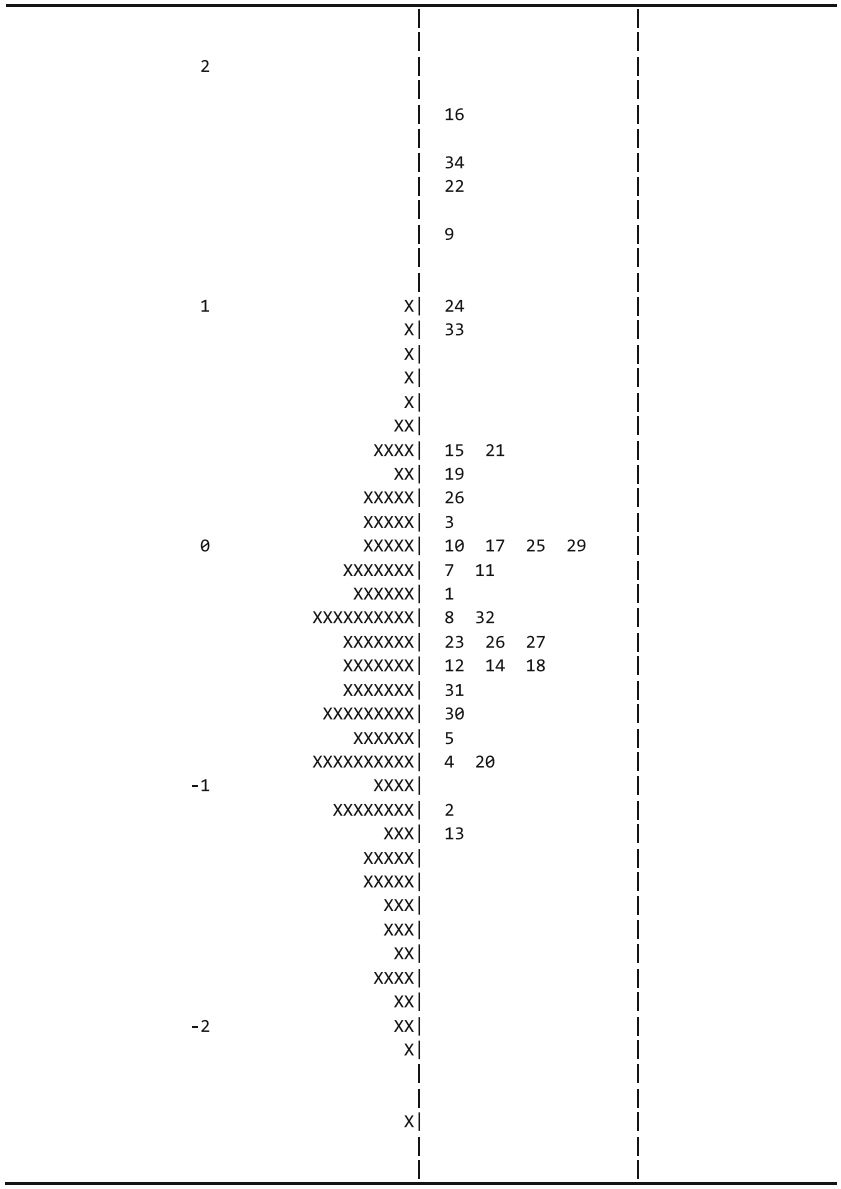

FIGURE 1 Map of person and item distribution. Each *X* represents 7.1 persons.

person) and item 34 (*hurt myself physically or taken dangerous risks with my health*) had correlations below .30. This indicates that these items contribute less to the overall total score in the sample used and that they do not discriminate well between participants with different levels of distress. It could be that these items are not suited to this particular population, that is, persons with mild to moderate symptoms of distress. This makes sense given three of these items measure self-harm, suicidal or violent behaviours. The endorsement proportions (Table 1) and skewness statistics also showed that very few clients endorsed the higher response categories, hence, these types of serious indications of psychological distress/impaired functioning may not apply to participants in the current study. The Rasch analyses of average item locations (item difficulty) are at least partly in support of this argument. According to these analyses, the items belonging to the Risk domain were the most difficult for the participants to endorse, meaning that the level of latent trait (amount of distress/psychological problems) required for endorsing these items is high. Relatedly, analysis of how the items contributed to measurement precision in certain parts of the scale showed a similar pattern. This analysis revealed that there was a relatively good match between clients' locations and items' locations in the middle of the latent trait continuum but precision was lower at the extremes. The test information at 0 logits was around 29. A test information of at least 10 (corresponding to a reliability of .90) was reached between −2.5 and +2.5 logits, which covered the great majority of the participants (see Figure 1). However, the scale showed less precision for this particular sample at lower levels of distress. Moreover, some items were located at trait levels above participants in the current sample but that poses less of a problem. It rather indicates that CORE-OM is suitable also for populations with more severe problems.

Previous studies have indicated that there are gender differences in some of the items of the CORE-OM, but it is unclear whether

differences are due to real differences between men compared to women or due to item bias. In the present study, the items in CORE-OM were examined for DIF with respect to gender. The results indicated DIF in some of the response categories for six of the items. This indicates that men and women at the same location of the latent trait have different probabilities of endorsing that response category. In general, the items displaying DIF required men to have a higher level on the latent trait to endorse the specific response category, except for item 11 (*tension and anxiety have prevented me from doing important things*). The items displaying DIF in our analysis are, however, not in line with previous research (e.g., Zeldovich et al., 2019). Only item 14 (*felt like crying*) mirrors findings from a previous study (Zeldovich et al., 2019) in that women had a higher probability of endorsing this item. The lack of overall alignment between our study and prior findings with respect to gender DIF could perhaps be due to cultural differences but needs to be further examined in future research. Furthermore, although one way to tackle DIF could be to include different norms for men and women (e.g., Zeldovich et al., 2019), a more reasonable and parsimonious solution is to revise the items to minimise gender bias. One concrete solution would be to reformulate item 14 to a less gender-sensitive option, such as *I have felt sad*, especially considering this item has displayed DIF also in another study (Zeldovich et al., 2019).

An important finding from the CTT-analyses as well as the Rasch analyses is that item 8 (*troubled by aches, pains or other physical problems*) was not a well-functioning item. It had poor item-total correlation, demonstrated local dependency with almost half of the other items and a poor fit to the Rasch model. This item has demonstrated poor fit also in other studies (e.g., Murray et al., 2014; Zeldovich et al., 2019) and as such, it appears to be problematic not only in the Swedish translation. Given the content of the item, comprehending general somatic complaints, it might align poorly with the other items of the CORE-OM and not adequately tap into psychological distress. This item is also formulated very broadly compared to the other items. One solution could be to separate it into two separate items, focusing on pains and aches on the one hand, and on other physical problems on the other hand.

Lastly, the functioning of the rating scale was examined by inspecting step and average measures for the response categories. Disordered step or average measures imply that higher categories do not represent/measure more of the construct than lower categories, and is something that can occur when the meaning of categories are ambiguous (i.e., 'Often' vs. 'Sometimes'; Smith et al., 2003). According to our analyses, only a couple of the items displayed disordered step or average measures (all belonging to the Risk domain). If the issue were ambiguous labelling of the response categories, disordered step or average measures would be apparent across all or most of the items, and not only in a few. Moreover, very few participants had chosen the higher response options (2–4) in these items, complicating interpretation of the findings. Therefore, the problem is likely not residing in the labels of response scale, but rather in the Risk items that appear relevant only for a minority of the participants in our sample. Other researchers have highlighted that the Risk items represent a psychometric problem (e.g., Zeldovich et al., 2019), and some scholars (e.g., Handscombe et al., 2016; Lyne et al., 2006) have suggested that they should be considered separately from the rest of the items. Our proposal is that when the CORE-OM is used in research, scholars should be wary of the psychometric issues of the Risk items and if distributions are skewed, collapsing the response categories into fewer categories could be a viable option. For clinical purposes, though, the Risk items are still of value, as a higher score on these items may represent very serious psychopathology.

4.1 | Strengths and limitations

The strengths of the current study include the use of a relatively large sample and the use of both CTT and IRT methods. The use of IRT to examine reliability of CORE-OM adds more nuanced information about the instrument's reliability, such that the instrument has its highest measurement precision in the middle of the latent trait but may less reliably capture very low or very high levels of distress. Moreover, the IRT analyses provided useful information about the match between item locations and clients' standing on the latent trait.

The major limitation of this study is that participants had relatively low levels of distress, hence, future research should include participants with more severe problems (i.e., psychiatric patients) to get a more complete picture of the measurement properties of CORE-OM across a wider range of distress. The sample also had an uneven gender distribution and the sample consisted of predominantly younger adults (mean age = 28 years). The results may thus be more generalizable to a younger population.

4.2 | Conclusions and suggestions for future research

In summary, the current study provided further evidence that Swedish translation of CORE-OM is a reliable instrument to measure psychological distress and the items were relatively well targeted to the sample of clients. Apart from that the items in the Risk domain were not well targeted to the participants in our sample and displayed disordered step or average measures, item 8 stands out as a particularly problematic item from a psychometric point of view and should either be removed or revised. Furthermore, analysis of DIF suggest that some of the items are not equally endorsed by men and women despite the same underlying trait level. Although our results are preliminary, the potential bias residing in these items should be examined further in future studies using representative samples, and if DIF is consistent, they might need to be revised to minimise gender bias. From a practitioner standpoint, it is also important to be aware that some items might not equally well capture distress in men compared to women.

Finally, an important application of IRT in clinical and outcome measurement is the possibility of examining longitudinal measurement invariance (Thomas, 2011). Longitudinal measurement invariance of

scale items means, for example, that changes in patients' scores after therapeutic intervention reflect true changes in the patients' standing on the latent variable and not variance in item parameters. Unfortunately, we were not able to examine this important aspect because our study only included cross-sectional data, but it is an important topic for future research on the CORE-OM.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data are available upon reasonable request.

ORCID

Elisabeth Åström https://orcid.org/0000-0003-2906-5409

REFERENCES

Barkham, M., Mellor-Clark, J., Connell, J., & Cahill, J. (2006). A CORE approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE system. *Counselling and Psychotherapy Research*, *6*(1), 3–15. https://doi.org/10.1080/14733140600581218

Bedford, A., Watson, R., Lyne, J., Davies, F., & Deary, I. J. (2010). Mokken scaling and principal components analyses of the CORE-OM in a large clinical sample. *Clinical Psychology & Psychotherapy*, *17*, 51–62. https://doi.org/10.1002/cpp.649

Bond, T. G., & Fox, C. M. (2004). *Applying the Rasch model*. Routledge, Taylor & Francis Group.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Derogatis, L. R., Lipman, R. S., & Covi, L. (1973). SCL-90: An outpatient psychiatric rating scale—Preliminary report. *Psychopharmacology Bulletin*, *9*(1), 13–28.

Derogatis, L. R., & Melisaratos, N. (1983). The brief symptom inventory: An introductory report. *Psychological Medicine*, *13*, 595–560. https://doi.org/10.1017/S0033291700048017

Elfström, M. L., Evans, C., Lundgren, J., Johansson, B., Hakeberg, M., & Carlsson, S. G. (2013). Validation of the Swedish version of the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM). *Clinical Psychology & Psychotherapy*, *20*(5), 447–455. https://doi.org/10.1002/cpp.1788

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardized brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, *180*(1), 51–60. https://doi.org/10.1192/bjp.180.1.51

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clinical outcomes in routine evaluation. *Journal of Mental Health*, *9*(3), 247–255. https://doi.org/10.1080/713680250

Evans, L. J., Beck, A., & Burdett, M. (2017). The effect of length, duration, and intensity of psychological therapy on CORE global distress scores. *Psychology and Psychotherapy: Theory, Research and Practice*, *90*(3), 389–400. https://doi.org/10.1111/papt.12120

Hambleton, R. K. (2005). Applications of item response theory to improve health outcomes assessment: Developing item banks, linking instruments, and computer-adaptive testing. In J. Lipscomb, C. C. Gotay, & C. Snyder (Eds.). *Outcomes assessment in cancer* (pp. 445–464). Cambridge University Press.

Hambleton, R. K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H. Tinsley & S. Brown (Eds.). *Handbook of applied multivariate statistics and modeling*. Academic Press. https://doi.org/10.1016/B978-012691360-6/50020-3

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Sage.

Handscombe, L., Hall, D. A., Hoare, D. J., & Shorter, G. W. (2016). Confirmatory factor analysis of Clinical Outcomes in Routine Evaluation (CORE-OM) used as a measure of emotional distress in people with tinnitus. *Health and Quality of Life Outcomes*, *14*(124), 1–9. https://doi.org/10.1186/s12955-016-0524-5

Holmqvist, R., Ström, T., & Foldemo, A. (2014). The effects of psychological treatment in primary care in Sweden—A practice-based study. *Nordic Journal of Psychiatry*, *68*, 204–212. https://doi.org/10.3109/08039488.2013.797023

Kean, J., Bordke, D. S., Biber, J., & Gross, P. (2018). An introduction to item response theory and Rasch Analysis of The Eating Assessment Tool (EAT-10). *Brain Impairments*, *19*(1), 91–102. https://doi.org/10.1017/BrImp.2017.31

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122.

Linacre, J. M. (2003). What is item response theory, IRT? A tentative taxonomy. *Rasch Measurement Transactions*, *17*(2), 926–927.

Lyne, K. J., Barrett, P., Evans, C., & Barkham, M. (2006). Dimensions of variation on the CORE-OM. *British Journal of Clinical Psychology*, *45*, 185–203. https://doi.org/10.1348/014466505x39106

Murray, A. L., McKenzie, K., Murray, K., & Richelieu, M. (2014). Mokken scales for testing both pre- and postintervention: An analysis of the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM) before and after counseling. *Psychological Assessment*, *26*(4), 1196–1204. https://doi.org/10.1037/pas0000015

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *Patient*, *7*(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

Palmieri, G., Evans, C., Hansen, V., Brancaleoni, G., Ferrari, S., Porcelli, P., Reitano, F., & Rigatelli, M. (2009). Validation of the Italian version of the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM). *Clinical Psychology & Psychotherapy*, *16*, 444–449. https://doi.org/10.1002/cpp.646

Penfield, R. D., & Gattamorta, K. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement*, *28*(1), 38–49. https://doi.org/10.1111/j.1745-3992.2009.01135.x

Skre, I., Friborg, O., Elgaroy, S., Evans, C., Myklebust, L. H., Lillevoll, K., Sørgaard, K., & Hansen, V. (2013). The factor structure and psychometric properties of the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM) in Norwegian clinical and non-clinical samples. *BMC Psychiatry*, *13*, 1–14, 99. https://doi.org/10.1186/1471-244X-13-99

Smith, E. V. Jr., Wakely, M. B., de Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement*, *63*(3), 369–391. https://doi.org/10.1177/0013164403063003002

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*(3), 291–307. https://doi.org/10.1177/1073191110374797

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442–1445. https://doi.org/10.1037/pas0000597

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer. https://doi.org/10.1007/978-1-4757-2691-6

Watson, R., van der Ark, L. A., Lin, L. C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in

clinical practice. *Journal of Clinical Nursing*, *21*(19–20), 2736–2746. https://doi.org/10.1111/j.1365-2702.2011.03893.x

Wilson, M. (2005). *Constructing measures. An item response modeling approach*. Erlbaum.

Wu, M., Adams, R., & Wilson, M. (1997). *ConQuest [computer program]*. Australian Council for Educational Research.

Zeldovich, M., & Alexandrowicz, R. W. (2019). Comparing outcomes: The clinical outcome in Routine Evaluation—Outcome Measure (CORE-OM) from an international point of view. *International Journal of Methods in Psychiatric Research*, *28*(3), 1–14. https://doi.org/10.1002/mpr.1774

Zeldovich, M., Ivanov, A. A., & Alexandrowicz, R. W. (2019). Dimensionality of the Russian CORE-OM from a Rasch perspective. *Journal of Applied Measurement*, *20*(3), 326–342. PMID: 31390606.

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, *67*(6), 361–370. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Åström, E., Sundström, A. E., & Lyrén, P.-E. (2023). Examining the psychometric properties of the Swedish version of Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM) in a clinical sample using classical test theory and item response theory. *Clinical Psychology & Psychotherapy*, *30*(2), 398–409. https://doi.org/10.1002/cpp.2808