Československá psychologie 2021 / ročník LXV / číslo 2
DOI: 10.51561/cspsych.65.2.163

PSYCHOMETRIC EVALUATION OF THE CLINICAL OUTCOME IN ROUTINE EVALUATION – GENERAL POPULATION: CZECH VERSION

ADAM KLOCEK, TOMÁŠ ŘIHÁČEK, HYNEK CÍGLER
Department of Psychology, Faculty of social studies, Masaryk University, Brno

ABSTRACT

Objectives. This study aimed to assess psychometric properties, such as reliability, construct validity, and cut-off scores, for the Czech version of the Clinical Outcome in Routine Evaluation – General Population (GP-CORE) questionnaire, a tool usable for repeated measurement of psychological distress within routine clinical settings.
Participants and setting. Two general populations and one clinical sample were used with N values of 420, 394, and 345, respectively.
Hypotheses. One of the competing theoretical factor solutions will demonstrate the best fit.
Statistical analysis. To examine the factor structure of the GP-CORE, a confirmatory multidimensional item response theory analysis (graded response model) was employed.
Results. The best fitting model was a bifactor solution representing one content domain of overall distress and two item wording domains (positively and negatively worded items). Clinical cut-off scores were determined to be 1.85 (men) and 1.90 (women).
Study limitations. The GP-CORE can be used as an unidimensional measure of overall distress, but users have to be aware of the influence of positive *vs*. negative item wording on the responses.

key words:
GP-CORE,
Clinical Outcome in Routine Evaluation – General Population,
psychological distress,
factor structure,
MIRT

klíčová slova:
GP-CORE,
Clinical Outcome in Routine Evaluation – obecná populace,
psychologický stres,
faktorová struktura,
multidimenzinální teorie odpovědi na položku

PROBLEM

The Clinical Outcome in Routine Evaluation adapted for the General Population (GP-CORE) is a measure derived from Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM), a self-report measure of general psychological distress (Barkham et al., 2001). The GP-CORE is focused primarily on the general population by not incorporating items regarding severe clinical problems. All risk and negatively worded high-intensity items were excluded from CORE-OM to reduce measurement bias within the general population. Two additional items were excluded for using the negatively worded phrase "my problems". The GP-CORE is a psychological distress (or reversed well-being) screening tool composed of 14 items (eight

Submitted: 5. 10. 2020; A. K., Department of Psychology, Faculty of Social Studies, Masaryk University, Joštova 10, Brno, 60200; e-mail: klocek.adam@mail.muni.cz
Funding: This study was funded by Specific University Research Grant No. MUNI/A/1071/2018 provided by the Czech Ministry of Education, Youth, and Sports.
The data that support the findings of this study are available from the corresponding author upon request.

positively and six negatively worded items, see Table 1). It was developed by Sinclair et al. (2005) and tested on a student sample in the *University Quality of Life and Learning Project*.

Due to the shift from the clinical to the general population, there is little reason to assume that the factor structures of the CORE-OM and the GP-CORE would be the same. A principal component analysis with oblique rotation resulted in a four-component structure: 1) six positively worded subjective well-being items, 2) four negatively worded subjective well-being items, 3) two social functioning items, and 4) two physical problems items. The first three components demonstrated modest to high associations; nevertheless, the correlation with the fourth component was negligible (Sinclair et al., 2005).

This four-component structure represents four main categories of the general quality of life (Wilkinson et al., 2012). Nevertheless, the authors believe that, conceptually, the GP-CORE can be used as a unitary measure. In existing studies, the GP-CORE is usually used on university students (e.g., Hammer & Vogel, 2013; Robinson et al., 2016), however, some studies sampled from Buddhist Vipassana retreat participants (Falkenström, 2010) or computer game players (Hagström & Kaldo, 2014). The GP-CORE authors called for psychometric validation studies among populations other than student populations, optimally among general population (Sinclair et al., 2005) and this call was not answered properly until the present study.

Psychometric and validation studies on GP-CORE are generally lacking. The four-component structure (positive, negative items, social functioning, physical symptoms) is problematic because it blends formal (i.e., positive *vs*. negative item wording) and content (i.e., subjective well-being, role functioning, physical symptoms) aspects together. Furthermore, as Schmitt and Stuits (1985) argue, a negative factor can emerge simply if 10 or more percent of respondents fill in the questionnaire with mixed item wording without sufficient attention. While the existing studies on the GP-CORE have not paid attention to this problem, item response theory (IRT) methods can be used to handle artificial factors, such as the mixed wording factor.

The GP-CORE repeatedly reaches good to excellent reliability in terms of internal consistency, from .85 (Cooke et al., 2004) to .90 (Richardson et al., 2017). Sinclair et al. (2005) also reported high test-retest reliability over one week with an *r* of .91. When computed separately for all four factors, Cronbach's α ranged from .75 to .85 (Mameli et al., 2018). Dividing the whole sample by gender, the GP-CORE reached higher internal consistency in female (α = .86) than male participants (α = .80) (Sinclair et al., 2005) and in Caucasians (α = .87) than Africans (α = .84) (Young & Campbell, 2014).

In terms of convergent validity, Sinclair et al. (2005) presented the relationship of the GP-CORE score with financial concerns (*r* = .25), sleeping difficulties (*r* = .27), and social support (*r* = -.32), as well as its correlations with other standard well-being measures: Beck Depression Inventory (BDI-I) (*r* = .77), BDI-II (*r* = .84), Brief Symptom Inventory (*r* = .75), Symptom Checklist (SCL-90-R) (*r* = .71), and CORE-NR (i.e., CORE-OM without risk items, *r* = .95) across samples (Sinclair et al., 2005).

Moreover, sex and age differences in GP-CORE mean scores were not significant (Sinclair et al., 2005; Cooke et al., 2006). The average item mean reported in Sinclair et al. (2005) was *M* = .99 (*SD* = .35) for the non-clinical student sample. Sinclair et al. (2005) created norms and clinical cut-off scores, establishing an average item mean exceeding 1.49 for men and 1.63 for women as indicators of elevated psychological distress.

Table 1 GP-CORE items (after Sinclair et al., 2005, modified)

ID	Item (Czech version)	Wording	Domain	PCA
1	I have felt tense anxious or nervous ("Měl/a jsem pocity napětí, strachu či nervozity.")	negative	Problems (anxiety)	2
2	I have felt I have someone to turn to when things go wrong ("Cítil/a jsem, že nemám někoho, na koho se mohu v případě potřeby obrátit.")	positive	Functioning (close)	3
3	I have felt OK about myself ("Byla/a jsem se sebou spokojený/á.")	positive	Subjective well-being	1
4	I have felt able to cope when things go wrong ("Měla/a jsem pocit, že zvládnu i těžké chvíle, kdyby měly přijít.")	positive	Functioning (general)	1
5	I have been troubled by aches, pains or other physical symptoms ("Trápily mě bolesti nebo jiné tělesné potíže.")	negative	Problems (physical)	4
6	I have been happy with the things I have done ("Byla/a jsem spokojen/a s tím, co jsem udělala/a.")	positive	Functioning (general)	1
7	I have had difficulty getting to sleep or staying asleep ("Měl/a jsem potíže usnout nebo jsem se předčasně budil/a.")	negative	Problems (physical)	4
8	I have felt warmth or affection for someone ("Cítila/a jsem k někomu opravdové přátelství nebo lásku (včetně rodiny).")	positive	Functioning (close)	3
9	I have been able to do most things I needed to ("Zvládl/a jsem většinu věcí, které jsem potřeboval/a udělat.")	positive	Functioning (general)	1
10	I have felt criticised by other people ("Měl/a jsem pocit, že mě druzí kritizují.")	negative	Functioning (social)	2
11	I have felt unhappy ("Cítil/a jsem se nešťastný/á.")	negative	Problems (depression)	2
12	I have been irritable when with other people ("V přítomnosti jiných lidí jsem byl/a podrážděný/á.")	negative	Functioning (social)	2
13	I have felt optimistic about my future ("Svou budoucnost jsem viděl/a optimisticky.")	positive	Subjective well-being	1
14	I have achieved the things I wanted to ("Dosáhl/a jsem toho, čeho jsem chtěl/a.")	positive	Functioning (general)	1

Note: The "Domain" column represents original CORE-OM domains (Barkham et al., 2001); the "PCA" column represents components from the principal components analysis reported by Sinclair et al. (2005).

Aim of the study

In the present study, we employed confirmatory IRT to evaluate the factor structure of the GP-CORE. Furthermore, we tested the measurement invariance between men and women, between younger and older adults, and between two different types of administration: standard administration and the GP-CORE framed with a retrospective instruction (see the Methods section for details). The following models were tested:

Model 1 was defined as unidimensional. In preceding studies, the GP-CORE has been used as a measure of overall distress in the general population by working with a sum score of all 14 items (e.g., Mamaeli et al., 2018; Robinson et al., 2016).

Model 2 was defined as two-dimensional, with factors composed of positively and negatively worded items. Whether these factors represent different kinds of overall distress or just a psychometric artefact is a matter of interpretation, as stated above (Schmitt & Stuits, 1985). The positive-wording latent factor was represented by Items 2, 3, 4, 6, 8, 9, 13, and 14, and the negative-wording latent factor was represented by Items 1, 5, 7, 10, 11, and 12.

Model 3 was defined as four-dimensional, following the original factor structure reported by Sinclair et al. (2005). The model was specified as follows: positively worded subjective well-being items (i.e., Items 3, 4, 6, 9, 13, and 14), negatively worded subjective well-being items (i.e., Items 1, 10, 11, and 12), social functioning (i.e., Items 2 and 8), and physical problems (i.e., Items 5 and 7).

Model 4 was defined as bifactor with one general factor explaining all 14 items and two specific latent factors representing positive and negative wording (the same as in *Model 2*). In this model, each item is explained by two factors, general (distress) and one of the specific wording factors. Therefore, this model represents both the formal and content aspects simultaneously.

METHODS

Participants

Power analysis about required sample size was not determined within this study because to test the models, we used archive data from three independent studies, namely, Pourová et al. (2019), Čevelíček et al. (2020), and Juhová et al. (2020). We focused on the general adult population. Informed consent was obtained from all individual participants included in the study.

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Descriptive characteristics are reported in Table 2.

Dataset 1 (hereafter referred to as "standard", Pourová et al., 2019) comprised data from Caucasian participants aged between 18 and 75 years from the Czech Republic who were given the standard administration of the GP-CORE. Eighteen percent of the participants suffered from serious chronic or somatic illness. Participants were recruited via social networks using snowball sampling. The goal of the original study was to validate the RNS-20 questionnaire, and the GP-CORE was used as one of the criterion measures.

Dataset 2 (hereafter referred to as "retrospective", Čevelíček et al., 2020) comprised data from Caucasian participants aged between 18 and 61 years from the Czech Republic. The GP-CORE was used in the context of an exploration of obstacles people may perceive when deciding whether to enter psychotherapy. Therefore, the sample comprised people who had considered beginning psychotherapy during the past five years; 48% of them started psychotherapy, while the rest decided not to. Participants were recruited via social networks using snowball sampling. The GP-CORE was framed with a retrospective instruction: "*Try to recall a period of time in which you were considering whether to begin psychotherapy. Please refer to an actual issue that bothered you in the past five years. Imagine that you would have been asked to*

Table 2 Descriptive characteristics of all three datasets

	Dataset 1 (standard)	Dataset 2 (retrospective)	Dataset 3 (embedded and clinical)
N	420	394	345
Gender (% women)	61.9%	80.7%	69.0%
Age	*M* = 30.7; *SD* = 11.1	*M* = 31.2; *SD* = 8.8	*M* = 37.5; *SD* = 11.9
Education	Primary (10.7%); High school (49.1%); University (40.2%)	Primary (0%); High school (34.0%); University (52.9%); Missing (13.1%)	Primary (5.6%); High school (57.6%); University (34.9%); Missing (1.9%)
Occupation	Unemployed or students or maternal leave (41.6%); Employed (45.6%); Entrepreneurs (9.8%) Missing (3%)	Unemployed or students or maternal leave (20.1%); Employed (63.1%); Entrepreneurs (8.6%); Missing (8.2%)	Unemployed or students or maternal leave or retired (47.4%); Employed/Entrepreneurs (46.4%) Missing (6.2%)

fill in this questionnaire at that time. Read the following instructions, please, and try to answer the questions as you would have at that time."

Dataset 3 (hereafter referred to as "clinical", Juhová et al., 2018) comprised data from Caucasian participants aged between 18 and 70 years from the Czech Republic who attended psychotherapy. This study focused on validation of the CORE-OM. Therefore, the participants were administered the full CORE-OM measure, and the GP-CORE items were extracted afterwards for the purpose of this study. This dataset was incorporated only for the computation of clinical cut-off scores.

Datasets 1 (standard) and 2 (retrospective) were merged for the purpose of analysis. From the whole sample of 823 participants, cases with missing items (n = 8) and age under 18 (n = 1) were excluded, resulting in a sample of N = 814 participants. The mean age was M = 31.16 (SD = 10.06, range from 18 to 75 years), and 71% of the sample were women.

Instrument

Clinical Outcome in Routine Evaluation for the General Population. The GP-CORE is composed of 14 items rated on a five-point Likert-type scale from 0 (not at all) to 4 (most or all of the time). Respondents are asked, "*How often did you experience the event described over the last week?*". Individual item scores can, therefore, range between 0 and 4 points (Sinclair et al., 2005), and the summative score of all 14 items can range between 0 and 56 points.

Statistical analysis

After data cleaning and preparation, statistical analyses were conducted using R software version 3. 5. 2 (R core Team, 2018). R packages: psych (Revelle, 2018) and mirt (Chalmers, 2012) were employed. Positively worded items (i.e., Items 2, 3, 4, 6, 8, 9, 13, and 14) were reversed before the analysis.

To compute clinical cut-off scores, we adopted Jacobson and Truax's (1991) formula (see Formula 1).

$$\frac{(Mean(clinical)*SD(normal)+Mean(normal)*SD(clinical)}{(SD(normal)+SD(clinical)} \quad (1)$$

To evaluate the reliability, McDonald´s omega (McDonald, 1999) was computed. To evaluate the factor validity, confirmatory factor analysis (CFA) is typically employed. One of the central assumptions of confirmatory factor analysis is the normality of the latent trait distribution. Violation of this assumption can lead to several parameter biases. In our study, we used IRT instead. In multidimensional IRT, and especially in the polytomous graded response model, parameter estimates are sufficiently robust (Wang et al., 2018). IRT is, therefore, more robust to non-normality elevated by categoric or ordinal items (e.g., Likert type) because it estimates item parameters directly instead of using covariances (Schulz, 2006; Reise et al., 1993). We believe that IRT provides more precise and detailed insight into the GP-CORE's underlying latent structure than classical test theory-driven CFA. Moreover, IRT showed better performance than CFA in terms of measurement invariance testing (cf. Meade & Lautenschlager, 2004).

Because we estimated multidimensional models, we adopted the multidimensional item response theory approach (De Ayala, 2013) by using the "mirt" package (Chalmers, 2012). We chose the graded response model (GRM, Samejima, 1969). Formula 2 represents the conditional probabilistic item response function of the GRM. The GRM estimates the probability of a participant's response to an item while overcoming the item's specific thresholds, as a sequential series of two-parameter models responsible for each threshold would (Reise et al., 1993).

Both item discrimination (a) and item location/difficulty (b) parameters varying among items were estimated. The discrimination parameter is similar to factor loading in the CFA approach, and the difficulty parameter is similar to an intercept in the CFA approach. However, in the IRT approach, the value of all four thresholds (b_1–b_4) between four pairs of neighbouring response categories are estimated, and for identification purposes. If the person's ability is equal to the threshold value b_j, the probability of observing any higher response option in such item is equal to the probability of observing any lower response category (Reise et al., 1993). These thresholds are estimated for each item separately, and thus, an item has five parameters in the GP-CORE (one slope and four thresholds). The probability of observing response option is given by Formula 2.

$$P(x = j \mid \theta) = \frac{1}{1+\exp[-a(\theta-b_{j-1})]} - \frac{1}{1+\exp[-a(\theta-b_j)]} = P^*_{j-1} - P^*_j \quad (2)$$

A confirmatory item response theory analysis was used to test several competing models of the GP-CORE's factor structure using Datasets 1 (standard) and 2 (retrospective). Marginal maximum likelihood estimation was used in models with up to two factors (i.e., *Models 1 and 2*). The quasi Monte-Carlo estimation method was used in models with more than two factors (i.e., *Models 3, 4, 5, and* 6) (Kuo & Kuyens, 2016). Covariances between latent variables were freely estimated (except for the bifactor model). If a latent variable was represented by only two items, these items' a-parameters were constrained to the same value. *Models 1* and *2* are nested in *Model 4* (bifactor) and, therefore, are directly comparable. *Model 4* employs item factor analysis adopted from Gibbons and Hedeker (1992) and Cai (2010), using "mirt: bfactor" syntax (Chalmers, 2012). The secondary or general latent factor (1x1 matrix)

is defined as being orthogonal to all specific latent factors with variance fixed to 1. All models were estimated on the merged dataset (Datasets 1 and 2).

Model fit was described using limited information C2 (Cai & Monro, 2014), which was used to estimate RMSEA, TLI, CFI, and SRMSR, with interpretation similar to a traditional CFA. We adopted the combination of Hu & Bentler's (1999) and Hooper et al.'s (2008) evaluation criteria: optimally, RMSEA should be below .05; however, values up to .10 are still considered a satisfactory fit. The SRMSR should not exceed .08. Optimally, TLI should be above .95; however, values above .90 are still considered a satisfactory fit.

To test the measurement invariance, we compared models using the likelihood ratio test, information criteria (BIC, saBIC), and delta fit index. Contrary to traditional CFA, metric and scalar levels of invariance are achieved simultaneously, constraining both slopes and thresholds (a, $b_{1–4}$) in IRT (Hui & Triandis, 1985). This results in the same item characteristic function across all the groups. The IRT invariance was tested across gender (i.e., male and female), age (i.e., younger and older cohort divided by median), and datasets (i.e., standard and retrospective administration). We considered two groups to be equivalent if the item parameters (a, b_1, b_2, b_3, b_4) were similar across groups. The chi-squared statistic is not considered a suitable indicator of model fit because it is usually significant when the dataset includes more than 400 cases. Therefore, we consider a model invariant across two groups if ΔCFI and ΔRMSEA do not exceed .01 (Kenny, 2011).

RESULTS

Reliability and validity

The internal consistency was satisfactory, with an omega of .90 (95% CI [.90, .92]). Internal consistency would be slightly higher if Item 5 was dropped (.91). Omega was also higher for women (.91) than for men (.90). However, the difference of .01 was negligible compared to the original study (Sinclair et al., 2005), where the difference between male and female reliabilities was approximately .06. Item mean was 1.9 (SD = .79). Within Dataset 1 (standard), reliability also reached satisfactory values with an omega of .87 (95% CI [.85, .87]). Within Dataset 2 (retrospective), reliability was similar, with an omega of .87 (95% CI [.83, .85]). The item mean for Dataset 1 (standard) was 1.6 (SD = .66). The item mean for Dataset 2 (retrospective) was 2.2 (SD = .68). The difference between the datasets was significant (Cohen's d = .895, $p < .001$).

Competing factor structures

Fit indices for all models are presented in Table 3. Summary statistics, item discrimination parameters, and item difficulty parameters for all models are reported in Table 4. The best fitting solution was the bifactor model (*Model 4*), although it did not differ significantly from *Model 2*. *Model 2*, consisting of positively and negatively worded factors, fitted the data significantly better than *Model 3*. *Model 2* did not fit the data significantly better than *Model 1*, although other fit indexes showed unambiguous support for *Model 2* over *Model 1*. Statistical comparisons using analysis of variance between the best fitting model (*Model 4*) and the rest of the models are presented in supplemental materials (Supplemental Table A), together with multidimensional discrimination and difficulty indices of the final bifactor model (Supplemental Table B).

Table 3 Fit indices for all models among all participants (N = 814)

	C2	Df	SRMSR	RMSEA (90% CI)	TLI	CFI	AIC	saBIC
Model 1 (*overall distress*)	825.15	77	.066	.110 [.103; .116]	.932	.943	29885.06	29991.91
Model 2 (*positive and negative wording*)	574.65	76	.057	.090 [.083; .097]	.955	.962	29717.31	29825.68
Model 3 (*four factors*)	2176.40	73	.321	.188 [.181; .195]	.801	.840	29916.55	30029.51
Model 4 (*bifactor*)	364.86	63	.045	.077 [.069; .084]	.967	.977	29539.21	29667.43

Note: Df = degrees of freedom; SRMSR = standardized root mean square residual; RMSEA = = root mean square error of approximation; TLI = Tucker-Lewis index; CFI = comparative fit index; AIC = Akaike information criterion; saBIC = sample adjusted Bayes information criterion; C2 = limited-information goodness of fit test statistic.

Final model

The best fitting model was the bifactor model (*Model 4*) with one general factor of overall distress (empirical r_{xx} = .90) and two wording factors responsible for positive (empirical r_{xx} = .53) and negative (empirical r_{xx} = .50) items. The bifactor structure provides a reasonable explanation and is well interpretable in terms of distinguishing the content and wording factors from each other. Additionally, given *Model 2* (two-dimensional wording factor structure), there is a high correlation between the latent factors of positive and negative wording (r = .89), establishing a good justification for considering higher-order or, in this case, a bifactor solution to better explain the data. Moreover, the bifactor model explained approximately 56% of the total variance (the G factor 45%, the positive wording dimension 5%, and the negative wording dimension 6%), while *Model 2* explained only 52% of the variance (positive wording 33% and negative wording 19%). Hence, we decided to continue with the standard bifactor solution (*Model 4*) as the final model (see Figure 1).

Measurement invariance

The fit indices for the configural, scalar, and latent factor means invariance between Datasets 1 and 2 are reported in Table 5. The bifactor model seems to be noninvariant on the scalar and factor means level. Significantly differential item functioning across both datasets within discrimination parameters was manifested by Items 3 ($p < .05$), 5 ($p < .001$), 6 ($p < .01$), and 13 ($p < .001$) by using all other items as anchors. Dataset 2 (retrospective) disposed of more elevated overall distress than Dataset 1 (standard). Although both datasets were sampled from the general, rather than clinical, population, participants in Dataset 2 (retrospective) may be more similar to the clinical population because they considered beginning psychotherapy. Because the datasets were not invariant, we decided to proceed with further invariance testing between genders and age groups only within Dataset 1 (standard). The bifactor structure (*Model 4*) held as the best fit scenario even after testing only in Dataset 1 (standard).

The fit indices for the configural, scalar, and latent factor means invariance between men and women are reported in Table 5. The bifactor model seems to be noninvariant

Table 4 Discrimination parameters and communality coefficients for Model 1, Model 2, Model 3, and Model 4, supplemented with correlations of latent variables, reliability coefficients, item means and standard deviations across the whole sample of 814 participants

Items	M/SD	Model 1		Model 2			Model 3					Model 4			
				Pos.	Neg.		PW	NW	SF	PP		G	PW	NW	
		a	h^2	*a*	*a*	h^2	*a*	*a*	*a*	*a*	h^2	*a*	*a*	*a*	h^2
I3	2.31/1.16	.88	.77	.88	–	.77	.79	–	–	–	.62	.88	.10	–	.78
I4	2.07/1.19	.80	.64	.81	–	.65	.69	–	–	–	.47	.80	.09	–	.66
I6	2.08/1.10	.85	72	.86	–	.74	.78	–	–	–	.61	.81	.29	–	.74
I9	1.77/1.14	.64	.41	.66	–	.43	.54	–	–	–	.29	.54	.56	–	.61
I13	2.19/1.26	.88	.77	.88	–	.78	.79	–	–	–	.62	.86	.20	–	.78
I14	2.13/1.12	.81	.65	.83	–	.68	.74	–	–	–	.54	.73	.54	–	.82
I1	2.39/1.12	.68	.46	–	.76	.58	–	.77	–	–	.59	.66	–	.47	.65
I10	1.70/1.20	.56	.31	–	.61	.38	–	.59	–	–	.35	.54	–	.29	.38
I11	2.17/1.35	.84	.71	–	.89	.79	–	.84	–	–	.70	.85	–	.21	.76
I12	1.82/1.14	.63	.40	–	.69	.47	–	.67	–	–	.45	.61	–	.33	.49
I2	1.75/1.31	.60	.36	.60	–	.35	–	–	.83	–	.70	.61	.04	–	.38
I8	1.02/1.15	.50	.25	.51	–	.26	–	–	.83	–	.70	.50	.12	–	.26
I5	1.71/1.27	.21	.05	–	.27	.07	–	–	–	.69	.48	.15	–	.45	.24
I7	1.93/1.38	.49	.24	–	.55	.30	–	–	–	.69	.48	.46	–	.38	.36
Cor F_1		–	–	–	.87	–	–	.59	.40	.29	–	–	–	0	–
Cor F_2		–	–	–	–	–	–	–	.25	68	–	–	–	–	–
Cor F_3		–	–	–	–	–	–	–	–	.19	–	–	–	–	–
Cor G		–	–	–	–	–	–	–	–	–	–	–	0	0	–
E.R_{xx}		.93		.92	.88		.90	.81	.72	.52		.90	.53	.50	

E. R_{xx} = empirical reliability; a = discrimination parameter; h_2 = communality; Items = item numbers according to position in the GP – CORE; Cor F_x = factor correlation; Pos. = positively worded items factor; Neg. = negatively worded items factor; PW = positive wellbeing factor; NW = negative wellbeing factor; SF = social functioning factor; PP = physical problems factor; G = general factor.

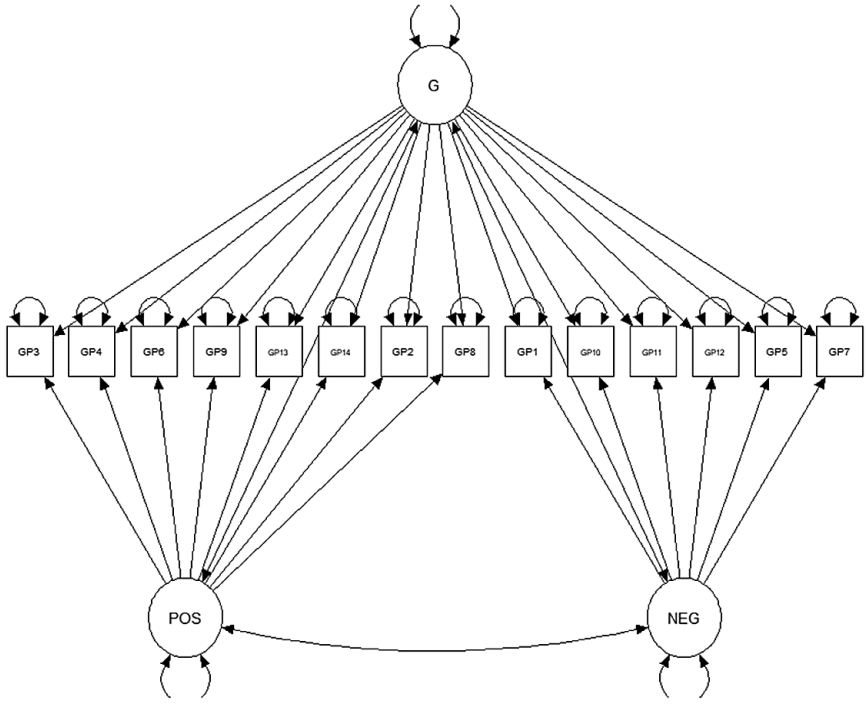

Figure 1 Bifactor model scheme with one general factor and two specific factors
Note: G = general factor in the bifactor solution; POS = positively worded items; NEG = negatively worded items

on the scalar and factor means level between men and women. Significantly differential item functioning across genders within discrimination parameters was manifested by Items 5 ($p < .05$), 6 ($p < .001$), 9 ($p < .05$), and 13 ($p < .05$) by using all other items as anchors. The bifactor model is at least partially invariant between gender groups if we set Item 6 free on the scalar (ΔCFI = .002) or factor means (ΔCFI = .003) level.

The fit indices for the configural, scalar, and latent factor means invariance between younger and older cohorts are reported in Table 5. The bifactor model seems to be noninvariant on the scalar and factor means level between younger and older participants. Significantly differential item functioning across genders within discrimination parameters was manifested by Items 2 ($p < .05$) and 7 ($p < .05$) by using all other items as anchors. The bifactor model is at least partially invariant between age groups if we set Items 2 and 7 free on the scalar (ΔCFI = .003) or factor means (ΔCFI = .002) level.

Clinical cut-off scores

To compute clinical cut-off scores, Dataset 3 (clinical), representing the clinical population, was compared to Dataset 1 (standard), representing the general population. Clinical cut-off scores computed according to Formula 1 yielded average item means exceeding 1.85 for men and 1.90 for women as indicators of potentially elevated distress; the overall cut-off score for a nondifferentiated sample was 1.89. The clinical population mean ($M = 2.19$, $SD = 1.02$) was greater than that of the non-clinical/general population ($M = 1.62$, SD = .89).

Table 5 Invariance indices of the bifactor model among all participants (n = 814)

Differences	Type of invariance	$\Delta\chi^2$	Δdf	CFI	RMSEA	saBIC	ΔCFI	ΔRMSEA	ΔsaBIC
Standard vs. Retrospective dataset	Configural			.971	.056	29583.8			
	Scalar	193.73*	82	.928	.068	29488.4	.043	.012	95.4
	Means	184.87*	1	.944	.060	29667.7	.016	.008	179.3
Dataset 1: Men vs. Women	Configural			.972	.057	15026.0			
	Scalar	178.55*	82	.948	.053	14969.5	.024	.004	56.5
	Means	.65	1	.947	.054	14967.3	.001	.001	2.2
Dataset 1: Younger vs. Older cohort	Configural			.970	.058	15042.8			
	Scalar	158.87*	82	.960	.052	14966.5	.010	.004	76.3
	Means	6.66*	1	.957	.054	14970.3	.003	.002	3.8

DISCUSSION

This study focused on the psychometric evaluation of the Clinical Outcome in Routine Evaluation – General Practice (GP-CORE). The Czech version of the GP-CORE questionnaire was successfully validated. Overall, the GP-CORE showed excellent reliability in the Czech sample. In terms of internal validity, we tested four different factor structure models. The internal consistency of a unidimensional model (*Model 1*) was sufficient, and its fit was borderline. Therefore, the use of the GP-CORE as a unidimensional measure is not disproved. However, administrators of the GP-CORE should be aware that participants tend to interpret well-being and distress as two partially separate constructs (see *Model 4*), even though, conceptually, they are facets of the same latent trait.

The four-factor structure demonstrated poor fit; therefore, Sinclair et al.'s (2005) model was not confirmed. Although it is plausible to suppose the existence of social functioning and physical symptoms dimensions, two items representing a whole factor might not function sufficiently since their covariance could be caused, for instance, by similar wording rather than a common latent variable.

A good fit of the two-dimensional model (*Model 2*) provided evidence that participants interpreted positively and negatively worded items differently. However, we do not assume that if participants were asked positively about overall distress, a different latent variable would emerge than if they were asked negatively. In fact, this negative-items factor could be a result of an inherent misinterpretation of reversed items by a non-negligible number of participants (Schmitt & Stuits, 1985).

The best factor structure solution was a combination of the unidimensional and the two-dimensional models. The bifactor model (*Model 4*) with a general factor representing overall distress and two specific factors accounting for positive and negative item wording yielded the best fit. In addition to the fact that factors within the two-dimensional model were highly correlated, the variance explained by the specific factors dropped to a minimum after we included the general factor. Additionally, our model explained a similar amount of variance (i.e., 56%) as the original four-factor model (i.e., 60%, Sinclair et al., 2005). On the one hand, Canivez (2016) considers the bifactor solution superior to other models for better interpretability of results. On the other hand, Murray and Johnson (2013) argue that there is a possible statistical bias causing bifactor models to demonstrate better fit over alternatives. Bifactor models might fit

better than other models simply because they successfully integrate potentially invalid responses (Reise et al., 2016). This information also argues for the preference of the bifactor model instead of the two-dimensional or unidimensional model.

The measurement invariance of the bifactor model (*Model 4*) between datasets, genders, and age groups was tested. Several GP-CORE items showed differences in functioning between the standard and retrospective modes of administration. This could be associated both with the different effects of each administration (standard *vs*. retrospective) and with inherent differences between the samples. Dataset 2 (retrospective) consisted of people who were considering starting psychotherapy and, therefore, had elevated levels of psychological distress. Gender and age group differences within Dataset 1 (standard) were also present only for several items. Although we identified the problematic items, the GP-CORE is already a short measure, and the removal of noninvariant items is not desirable.

Despite the existence of noninvariant items, the bifactor model was acceptable without modifications. We visually checked the Czech wording of items to see whether the existence of the specific factor may be explained by similarities in wording, but we were unable to observe such similarities. Indeed, the existence of the specific factors seems to be driven by the negative wording *per se*.

Strengths and Limitations

To our knowledge, our study was the first to assess the GP-CORE factor structure using the IRT methodology and a confirmatory, rather than exploratory, approach. Furthermore, this study was the first to validate the GP-CORE on a sample of the general population (i.e., beyond university students).

The study has several limitations associated with the samples used. First, the results are generalizable primarily for women (71% of the total sample) and younger people (although the age range was between 18 and 75, the majority of the sample was approximately 30 years old). Second, Datasets 1 (standard) and 2 (retrospective) were obtained via snowball sampling using social networks. These sampling procedures are considered to lead to less representative samples than other procedures. Furthermore, Dataset 2 (retrospective) is biased towards the clinical population. Nearly half of the participants had experienced psychotherapy in the past. Third, the sample size of Dataset 1 (standard) alone was relatively small/borderline to reliably test the measurement invariance. It should also be noted that the Czech version of the GP-CORE was used in this study, and the generalizability of the results to other language versions must be verified in future studies.

CONCLUSION

The aim of the study was to test the factorial structure of the GP-CORE and its measurement invariance. We found that although the measure *can* be used as an essentially unidimensional measure of psychological distress/well-being, the structure is better explained by a bifactor model that takes into account the positive/negative wording of items. We assume that the overall distress is manifested through participants' responses to all 14 items (the general factor in the bifactor model), while item wording seems to interfere with this overall distress manifestation.

REFERENCES

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., … McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and*

Clinical Psychology, 69(2), 184. https://doi.org/10.1037//0022-006x.69.2.184

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581-612. https://doi.org/10.1007/s11336-010-9178-0

Cai, L., & Monro, S. (2014). *A new statistic for evaluating item response theory models for ordinal data.* Technical Report. National Center for Research on Evaluation, Standards, & Student Testing.

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for multidimensionality and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 247-271). Gottingen, Germany: Hogrefe.

Čevelíček, M., Tarinová, A., & Řiháček, T. (2020). Vnímané překážky při vstupu do psychoterapie [Perceived barriers to entering psychotherapy]. *Československá psychologie, 65*(1), 1-13.

Cooke, R., Barkham, M., Audin, K., Bradley, M., & Davy, J. (2004). Student debt and its relation to student mental health. *Journal of Further and Higher Education*, *28*(1), 53-66. https://doirg/10.1080/030987703200016181 4

Cooke, R., Bewick, B. M., Barkham, M., Bradley, M., & Audin, K. (2006). Measuring, monitoring and managing the psychological well-being of first year university students. *British Journal of Guidance & Counselling*, *34*(4), 505-517. https://doi.org/10.1080/03069880600942624

Chalmers, R., P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

De Ayala, R. J. (2013). *The theory and practice of item response theory.* Guilford Publications.

Falkenström, F. (2010). Studying mindfulness in experienced meditators: A quasi-experimental approach. *Personality and Individual Differences*, *48*(3), 305-310. https://doi.org/doi:10.1016/j.paid.2009.10.022

Gibbons, R., & Hedeker, D. (1992). Full-information item factor analysis. *Psychometrika*, *57*, 423-436. https://doi.org/10.1177/014662168801200305

Hagström, D., & Kaldo, V. (2014). Escapism among players of MMORPGs—conceptual clarification, its relation to mental health factors, and development of a new measure. *Cyberpsychology, Behavior, and Social Networking*, *17*(1), 19-25. https://doi.org/10.1089/cyber.2012.0222

Hammer, J. H., & Vogel, D. L. (2013). Assessing the utility of the willingness/prototype model in predicting help-seeking decisions. *Journal of Counseling Psychology, 60*(1), 83. https://doi.org/10.1037/a0030449

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60. https://doi.org/10.21427/D7CF7R

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131-152. https://doi.org/10.1177/0022002185016002001

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12. https://doi.org/10.1037//0022-006x.59.1.12

Juhová, D., Řiháček, T., Cígler, H., Dubovská, E., Saic, M., Černý, M., Dufek, J., & Evans, Ch. (2018). Česká adaptace dotazníku CORE-OM: vybrané psychometrické charakteristiky [Czech adaptation of the CORE-OM: Selected psychometric properties]. *Československá psychologie, 62*(1), 59-74.

Kenny, D. A. (2011). Measuring model fit. Retrieved from http://davidakenny.net/cm/fit.htm

Kuo, F. Y., & Nuyens, D. (2016). A practical guide to quasi-Monte Carlo methods. https://people.cs.kuleuven.be/~dirk.nuyens/taiwan/QMC-practical-guide-20161107-1up.pdf

Mameli, C., Biolcati, R., Passini, S., & Mancini, G. (2018). School context and subjective distress: The influence of teacher justice and school-specific well-being on adolescents' psychological health. *School Psychology International*, *39*(5), 526-542. https://doi.org/10.1177/0143034318794226

McDonald, R. P. (1999). *Test theory: A unified approach*. Lawrence Erlbaum Associates Publishers.

Meade, A. W., & Lautenschlager, G. J. (2004). *Same Question, Different Answers: CFA and Two IRT Approaches to Measurement Invariance.* Symposium presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence, 41*(5), 407-422. https://doi.org/10.1016/j.intell.2013.06.004

Pourová, M., Řiháček, T., & Žvelc, G. (2020). Validation of the Czech Version of the Relational Needs Satisfaction Scale. *Frontiers in*

Psychology, 11, 359. https://doi.org/10.3389/fpsyg.2020.00359.
R Core Team (2018) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org
Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51*(6), 818-838. https://doi.org/10.1080/00273171.2016.1243461
Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological bulletin, 114*(3), 552. https://doi.org/10.1037/00332909.114.3.552
Revelle, W. (2018) psych: Procedures for personality and psychological research, Evanston, Illinois, USA: Northwestern University. https://CRAN.R-project.org/package=psychVersion=1.8.12
Richardson, T., Elliott, P., Roberts, R., & Jansen, M. (2017). A longitudinal study of financial difficulties and mental health in a national sample of British undergraduate students. *Community Mental Health Journal*, *53*(3), 344-352. https://doi.org/10.1007/s10597-016-0052-0
Robinson, A. M., Jubenville, T. M., Renny, K., & Cairns, S. L. (2016). Academic and mental health needs of students on a Canadian campus. *Canadian Journal of Counselling and Psychotherapy*, *50*(2). Retrieved from https://cjc-rcc.ucalgary.ca/article/view/61100
Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. http://www.psychometrika.org/journal/online/MN17.pdf.
Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*(4), 367-373. http://dx.doi.org/10.1177/014662168500900405.
Schulz, W. (2006). Testing parameter invariance for questionnaire indices using confirmatory factor analysis and item response theory. Advance online publication. Paper prepared for the Annual Meeting of the American Educational Research Association in San Francisco, April 7–11, 2005.
Sinclair, A., Barkham, M., Evans, C., Connell, J. & Audin, K. (2005). Rationale and development of a general population well-being measure: psychometric status of the GP-CORE in a student sample. *British Journal of Guidance and Counselling, 33*, 153-173. https://doi.org/10.1080/03069880500132581
Wang, C., Su, S., & Weiss, D. J. (2018). Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivariate Behavioral Research, 53*(3), 403-418. https://doi.org/10.1080/00273171.2018.1455572
Wilkinson, S., Mistral, W., Griffin, C., Parsons, J., & Williamson, V. (2012). University Students' Wellbeing, Alcohol and Drug Use (USWAD): A pilot study: Final Draft Report. University of Bath (N.d.), 1-43. http://www.academia.edu/download/41769451/USWAD_FINAL_Report_September_2012.docx
Young, C., & Campbell, M. (2014). Student wellbeing at a university in post-apartheid South Africa: A comparison with a British university sample using the GP-CORE measure. *British Journal of Guidance & Counselling*, *42*(4), 359-371. https://doi.org/10.1080/03069885.2013.779638

SOUHRN

Cíle. Tato studie si kladla za cíl zhodnotit psychometrické vlastnosti české verze škály Clinical Outcome in Routine Evaluation - General Population (GP-CORE), konkrétně poskytnout informace o reliabilitě, konstruktové validitě a klinickém cut-off skóru. GP-CORE je nástroj využitelný pro opakované měření psychologického stresu.
Vzorek a design. Byly využity tři vzorky respondentů, z nichž dva pocházely z obecné (N = 420 a 394) a jeden z klinické populace (N = 345).
Hypotézy. Jedno z faktorových řešení bude mít superiorní fit oproti ostatním.
Statistické analýzy. K ověření faktorové struktury GP-CORE bylo využito konfirmační multidimenzionální teorie odpovědi na položku (graded response model).
Výsledky. Jako finální model byl zvolen bi-faktorový model reprezentující obecný obsahový faktor distresu a dva metodické faktory způsobené pozitivní a negativní formulací položek. Kromě toho, že vykazoval nejlepší fit, byl také dobře teoreticky interpretovatelný. Klinický cut-off skór pro průměrnou hodnotu napříč položkami byl ustanoven na 1,85 pro muže a 1,90 pro ženy.
Limity. GP-CORE lze využít jako jednodimenzionální nástroj pro měření obecného psychologického stresu, ale uživatelé by si měli být vědomi také možného vlivu pozitivní či negativní formulace položek na odpovědi respondentů.

SUPPLEMENT

Supplemental Table A ANOVA comparison of $\Delta\chi^2$ between given models among all 814 participants

Model (a)		Model (b)	$\Delta\chi^2$
Model 4	>	Model 1	$\chi 2(14) = 373.85$
Model 4	>	Model 2	$\chi 2(13) = 204.10$
Model 4	>	Model 3	$\chi 2(10) = 397.34$

Within the final bifactor model, according to a multidimensional discrimination index (*a parameter* for the multidimensional latent space) and multidimensional difficulty indexes (b_j *parameters* for the multidimensional latent space) (see Table B), the smallest amount of information about the combination of participants' responses to positively and negatively worded psychological distress and general factor of overall distress free from this dichotomy provides the item 5, whereas the largest amount of information provides item 14. By visual evaluation of the item-plots we inferred that a positive wording dimension brings more information about overall distress than negative dimension.

Item 5 is characterized by the lowest multidimensional discrimination parameter, however the model fit increases only a little when this item is omitted. The model fit could be increased dramatically if we omit item 8: "*I have felt warmth or affection for someone*." Yet, GP-CORE was created as a screening tool for several unique distressing elements. Every item covers a specific problem that should be screened among general population as a checklist in order to identify people who might be in greater distress, and who might need a further help. Omitting problematic items could be only one of the solutions. Reise et al. (1993) shows that leaving items with differential items functioning dependent on group membership within the questionnaire need not result into a bias in estimation. However, if the goal is to maintain the original four-factor structure, each factor needs to be represented by more items, particularly to the social functioning and physical problems (represented only by two items so far). Unfortunately, these new items would need a novel standardization and validation study, confirming the four-factor structure.

Supplemental Table B Multidimensional discrimination and difficulty indexes of all four thresholds across all 14 GP-CORE items (n = 814)

Item	Multidim. discrimination index	Multidim. difficulty index Threshold 4	Multidim. difficulty index Threshold 3	Multidim. difficulty index Threshold 2	Multidim. difficulty index Threshold 1
I1	2.32	-2.01	-.98	-.02	1.17
I2	1.32	-1.20	-.20	.81	1.98
I3	3.19	-1.84	-.78	.22	.94
I4	2.34	-1.70	-.49	.44	1.30
I5	.96	-1.56	-.19	1.01	2.78
I6	2.86	-1.77	-.66	.49	1.44
I7	1.27	-1.31	-.44	.44	1.73
I8	1.01	-.31	.99	2.18	3.53
I9	2.11	-1.56	-.12	.82	1.73
I10	1.33	-1.53	-.15	.91	2.31
I11	3.06	-1.28	-.46	.16	.92
I12	1.65	-1.64	-.35	.74	2.12
I13	3.17	-1.48	-.56	.26	.96
I14	3.64	-1.66	-.62	.44	1.18