Sočan G, De Boer D, Murko E, Kralj M, Ropret N, Zaletel M. Psychometric validation of an instrument for measuring patient experiences with outpatient healthcare.
Zdr Varst. 2023;62(3):153-161. doi: 10.2478/sjph-2023-0021.

PSYCHOMETRIC VALIDATION OF AN INSTRUMENT FOR MEASURING PATIENT EXPERIENCES WITH OUTPATIENT HEALTHCARE

PSIHOMETRIČNA VALIDACIJA INŠTRUMENTA ZA MERJENJE IZKUŠENJ PACIENTOV Z ZUNAJBOLNIŠNIČNO ZDRAVSTVENO OBRAVNAVO

Gregor SOČAN [1*], Dolf DE BOER [2], Eva MURKO [3], Marcel KRALJ [3], Nina ROPRET [3], Metka ZALETEL [3]

[1] University of Ljubljana, Department of Psychology, Aškerčeva cesta 2, 1000 Ljubljana, Slovenia
[2] Netherlands Institute for Health Services Research, Otterstraat 118, 3513 CR Utrecht, The Netherlands
[3] National Institute of Public Health, Trubarjeva cesta 2, 1000 Ljubljana, Slovenia

Received: Sep 6, 2022
Accepted: May 11, 2023

Original scientific article

ABSTRACT

Keywords:
Patient reported experience measures
Reliability
Validity
Factor analysis
Slovenia

Aim: Recently, a patient-reported experience measure (PREM) was developed in Slovenia to assess patients' experiences with outpatient specialist healthcare clinics. The aim of this study was to evaluate the psychometric properties (including factor structure, reliability, convergent validity, and response distribution) of the questionnaire.

Methods: The sample consisted of 8,406 adult participants treated in 171 specialist clinics from different medical fields. Participants voluntarily and anonymously responded to either the paper or online survey.

Results: Descriptive statistics show meaningful response patterns with a general tendency towards favourable assessments. The psychometric analyses of the scales evaluating doctor's and nurses' work, respectively, generally showed a good fit of the unidimensional factor model as well as the Rasch model, high factor loadings and very good to satisfactory reliability. The Rasch scaling showed that these scales were most informative for patients with relatively unfavourable experience ratings.

Conclusions: The results are similar to those found in previous evaluations of PREMs in other countries. Given its good psychometric properties, the Slovenian PREM can be recommended for healthcare evaluations in Slovenia and as a model for the development of similar PREMs in other countries.

IZVLEČEK

Ključne besede:
merjenje izkušenj pacientov
zanesljivost
veljavnost
faktorska analiza
Slovenija

Namen: *V Sloveniji je bil pred kratkim razvit Vprašalnik o izkušnjah pacientov z obravnavo v specialističnih ambulantah. Namen te raziskave je bil preveriti njegove psihometrične lastnosti (vključno s faktorsko strukturo, zanesljivostjo, konvergentno veljavnostjo in porazdelitvami odgovorov).*

Metode: *V raziskavi je sodelovalo 8.406 odraslih udeležencev, ki so bili obravnavani v 171 specialističnih ambulantah različnih zdravstvenih dejavnosti. Udeležba je bila prostovoljna in anonimna. Udeleženci so lahko izbirali med papirnato in spletno verzijo vprašalnika.*

Rezultati: *Opisne statistike kažejo smiselne vzorce odgovorov, pri čemer so se nizke ocene v splošnem pojavljale razmeroma redko. Psihometrične analize lestvic zdravnikovega dela in dela sester so v splošnem pokazale dobro prileganje enodimenzionalnega faktorskega in Raschevega modela, visoke faktorske uteži in zelo dobro (pri delu zdravnika) oz. zadovoljivo (pri delu sester) zanesljivost dosežka na lestvici. Rezultati Raschevega lestvičenja so pokazali, da sta ti lestvici najbolj informativni za paciente z relativno nizkimi ocenami svoje izkušnje.*

Zaključek: *Rezultati so podobni tistim, ki so jih pokazale analize podobnih vprašalnikov v drugih državah. Glede na dobre psihometrične lastnosti lahko slovenski vprašalnik priporočimo za uporabo tako pri ocenjevanju zdravstvenih storitev v Sloveniji kot tudi kot model za razvoj podobnih vprašalnikov v primerljivih državah.*

***Correspondence:** gregor.socan@ff.uni-lj.si

1 INTRODUCTION

The concept of patient centeredness (PC) has been widely discussed in recent years (1-3). Research has shown that PC is positively related to patient satisfaction, well-being, and self-management, which are especially relevant in chronic disease management. Patient Reported Experience Measures (PREMs) can effectively assess quality of care and patient-centredness (4). Patient reported experience measures are tools that capture "what" happened during an episode of care, and "how" it happened from the patient's perspective (5).

The use of PREMs is also recommended by the Organization for Economic Cooperation and Development (OECD) (6) because of the widely recognized relationships among patient experience, the process of care, and health outcomes (7). At the national level, in almost all OECD countries PREMs are collected through surveys, covering population samples of patients receiving inpatient or outpatient care (6). The OECD has developed internationally comparable patient experience indicators related to access to healthcare, autonomy in care and treatment decisions and communication with the physician during ambulatory care due to their relevance and importance across health systems (8). Otherwise, an internationally accepted standardized PREM tool for ambulatory or hospital healthcare for Europe or worldwide does not exist. Different countries use their own standardized and validated instruments. For example, Norway uses the Norwegian Generic questionnaire about experiences and importance (9) and the outpatient experiences questionnaire, OPEQ (10). The USA uses the American Consumer assessment of healthcare providers and systems (11), while the Netherlands has the Dutch PREMs Consumer Quality Index questionnaires for inpatient and outpatient hospital care (12). In an Australian systematic review, a total of 88 PREMs were identified. More than one-third of these instruments were designed for inpatient care services, roughly 25% percent for primary care services and only 12.5% for outpatient care services. Over 40% of these 88 PREMs were developed and tested in languages other than English (5).

The Slovenian National Institute of Public Health (NIJZ) led a project on PREMs and patient reported outcome measures (PROMs) between 2017 and 2019. During the project the first national questionnaire on patient experiences with outpatient specialist healthcare clinics was developed, which was tailored and adapted to the Slovenian healthcare system. To enable international comparisons, four of the OECD quality indicators on patient experiences were included:1) the doctor spending enough time with patients during a consultation; 2) the doctor providing easy-to-understand explanations; 3) the doctor providing the opportunity to ask questions and express concerns; and 4) the doctor involving patients in decisions that affect their care and treatment (6). Twenty-one OECD countries, including Slovenia, provide data on the person-centeredness of ambulatory care, and the data are annually published in the OECD Health at a Glance report as indicators of quality of care (13).

Between 2006 and 2012, the Slovenian Ministry of Health conducted a national patient experience survey in acute and psychiatric hospitals, although the survey was then discontinued (14). When developing a new PREM tool for specialist ambulatory healthcare, we thus built on this earlier experience.

The aim of the present study was to validate the outpatient PREM instrument with the objective of evaluating some of its psychometric properties in adult Slovenian patients receiving treatment in outpatient specialist healthcare clinics.

2 METHODS

2.1 Development of the questionnaire

The questionnaire was developed in several different stages: A literature review, expert evaluations and discussions with Steering Committee members, focus groups with patients and cognitive interviews were conducted to identify key aspects of healthcare from both provider and patient perspectives, and to test patients' understanding of the items (15). The Anglo-American and Scandinavian literature was searched for aspects of patient experiences of relevance in an outpatient setting, such as admission, privacy, work of doctors, work of nurses, information about patient rights and continuation of treatment and overall score (9-12, 16). The questionnaire was piloted in three different settings after which further changes were made to improve the content validity of the items.

2.2 Data collection

The study was conducted from June 1 to June 30, 2020, as a cross-sectional survey using a paper or online questionnaire. The patients surveyed received specialist healthcare in 171 specialist clinics in the following medical fields: neurology and neurosurgery, cardiology, ophthalmology, gynaecology and obstetrics, and orthopaedics. At the end of the specialist visit, the nurse handed each patient a letter with a paper questionnaire and a prepaid envelope and invited them to participate in research. The patient also had the option of completing an online survey, for which they received a one-time password. Out of the total 8,406 participants, 7,877 (94%) completed the paper survey, and 529 (6%) completed the online survey. Patients were also given information about the study, including a statement that participation was voluntary and anonymous, and that their responses would never be linked to them as individuals. Participation or nonparticipation in the study did not affect patient's treatment or healthcare. No reminders were sent.

2.3 Participants

Of the 8,406 patients who participated in the study, 68.0% were female and 32.0% were male. The two largest age groups were those aged 65 to 74 years (25.0% of all respondents) and 55 to 64 years (21.1%). More than 60.0% of participating patients had completed high school or above. A total of 53.8% of respondents rated their own health as good or very good, and 7.8% rated it as poor or very poor, while the rest of the patients rated their health as average. A total of 45.1% of the patients surveyed had a chronic disease. The response rate was estimated at roughly 31%. A comparison of the sociodemographic characteristics of the respondents and all patients having an eReferral for a specialist clinic for the time the survey was conducted, based on the eReferral database, revealed the predominance of female and relatively young respondents (further details are presented in the online supplement: https://osf.io/4guch/?view_only=1c3474c36e95481b959dd22a2cd03460).

2.4 Data analysis

Returned paper questionnaires were manually entered into the database via the input mask. To check the quality of data entry, 10% of all questionnaires were double-entered. In the case of an online survey, the data was recorded in the database automatically. Forty-five records with a completion-rate less than 50% were excluded from the analysis.

Patient experience is not a unitary construct, so we tailored the analyses according to the characteristics and content of the items. The items that did not rate the patient experience were not relevant to the analysis. These included:

- questions asking for general information, that did not assess patient's experience, including age, education, presence of chronic disease, reasons for visiting the doctor etc. (questions Q1, Q2, Q31-Q34), and
- open-ended questions (questions Q4, Q28, and Q29) that asked for specific additional comments.

The questions related to the doctor's (Q12-Q17) and nurses' work (Q21-Q23) were analysed as homogeneous psychometric scales. We used confirmatory factor analysis with the WLSMV estimator with pairwise estimation to assess the dimensionality of the scale and the discrimination power of the items. The "not applicable" (NA) responses were treated as missing values in these analyses. Item Q16 is a group of four separate ratings, so they were analysed as separate indicators. Because they are all related to the doctor's explanations, we allowed correlated residuals between these four items. For each scale we computed two measures of score accuracy: the coefficient omega, indicating the proportion of the scale score variance, attributable to the common factor, and the coefficient alpha, a standard lower bound to the scale score reliability. After the factor analysis, we performed a Rasch analysis, based on the partial credit model for ordinal responses. In addition, we applied a unidimensional variant of the optimal scaling, called correlational aspects optimization. In this analysis, each response category is assigned a scale value so that the weighted sum of the items is as unidimensional as possible.

To assess the fit of the factor analysis model, we used the criteria proposed by Hu and Bentler (17): RMSEA<.06, CFI>.95, SRMR<.08. For the Rasch analysis, we used the criterion value of 1.3 for the INFIT and OUTFIT statistics (18).

The significance level of α=5% was used for all statistical tests.

For the remaining items, standard psychometric analysis was not meaningful, either because they were standalone measures (for instance, a general evaluation in item Q27), or because an item was presented conditionally on the response to the preceding items (for instance, the waiting time and receiving explanations for a prolonged waiting time in items Q7 and Q8). The frequency distributions of the responses were calculated for these items, and their validity is based on their content. In examining the frequency distributions, we checked that the proportions of missing values and NA responses were reasonable (it should be noted that such an evaluation depends on the item content; for instance, a very low proportion of NA responses should be expected for item Q12, but less so for item Q14), and that no unusual frequency patterns emerged.

The main psychometric analyses (factor analysis, coefficient omega, and Rasch scaling) allow the presence of missing values. Only for the computation of the coefficient alpha and for the optimal scaling did we impute the missing values using the random forest algorithm.
All analyses were performed in R (19). We used the packages lavaan (20) for the confirmatory factor analysis, MBESS (21) for the calculation of coefficient omega, eRm (22) for the Rasch analysis, aspect (23) for optimal scaling, psych (24) for general descriptive statistics, and missForest (25) for missing data imputation.

3 RESULTS

Table 1 shows the percentages of responses to the questionnaire items, except for the open-ended and non-evaluation items. In general, the proportion of unfavourable evaluations (typically these were the "not at all" and "not particularly" responses) was low in most cases. The proportion of "not applicable" responses varied widely across items. For items evaluating the doctor's work, it ranged from about one-quarter to one-third of responses. In contrast, this proportion was much lower for most items related to the reception process, privacy, and the nurses' work.

Table 1. Frequencies and percentages of item responses.

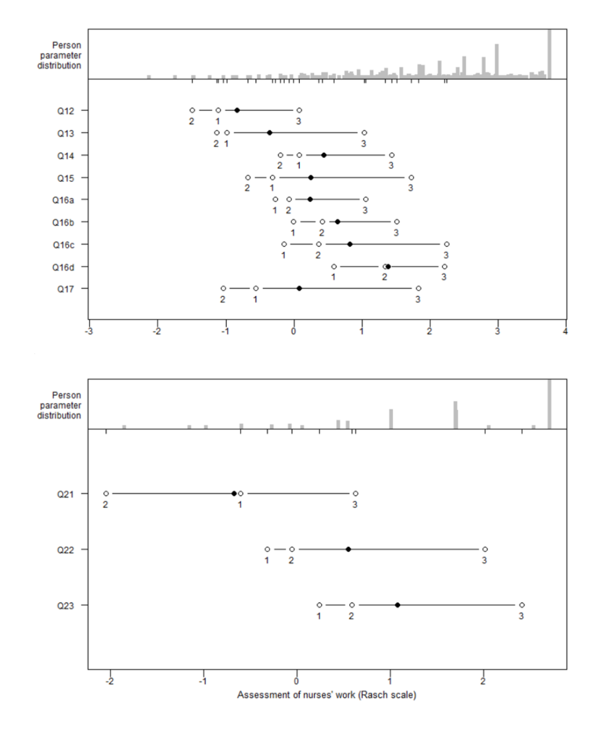
Item*	Responses (%)					
	Yes	No				NA
Q3 Problems making appointment	240 (3.0)	7231 (90.1)				552 (6.9)
	No problems	Minor problems	Major problem			NA
Q5 Waiting list problem	6422 (77.3)	1235 (14.9)	355 (4.3)			296 (3.6)
	Not at all	Not particularly	Yes, to some extent	Yes, certainly		NA
Q6 Kind reception	33 (0.4)	33 (0.4)	395 (4.7)	7812 (93.6)		73 (0.9)
	Less than 15 min	15-30 min	30-60 min	More than 60 min		
Q7 Waiting beyond scheduled time	5045 (60.5)	2468 (29.6)	655 (7.9)	175 (2.1)		
	Yes	No				NA
Q8 Needed to wait for Information	765 (12.5)	860 (14.1)				4476 (73.4)
	Yes	Partly	No			
Q9 Premises in order	8028 (96.3)	267 (3.2)	42 (0.5)			
	Not at all	Not particularly	Yes, to some extent	Yes, certainly		NA
Q10 Sufficient privacy	39 (0.5)	53 (0.7)	688 (8.5)	7326 (90.4)		
Q12 Doctor polite and respectful	21 (0.3)	42 (0.5)	425 (5.1)	7830 (94.1)		
Q13 Explanation easy to understand	31 (0.4)	78 (0.9)	746 (9.0)	6936 (83.3)		538 (6.5)
Q14 Could ask doctor questions	99 (1.2)	141 (1.7)	716 (8.6)	5159 (62.2)		2183 (26.3)
	Not at all	Not particularly	Yes, to some extent	Yes, certainly	Did not want	No need
Q15 Involved in decisions	62 (0.8)	108 (1.3)	840 (10.2)	4907 (59.4)	171 (2.1)	2175 (26.3)
Did the doctor give an explanation…?	Not at all	Not particularly	Yes, to some extent	Yes, certainly		NA
Q16a …the necessity of the intervention?	65 (0.9)	119 (1.6)	491 (6.6)	4596 (62.2)		2124 (28.7)
Q16b … how it will proceed?	96 (1.4)	174 (2.5)	556 (8.1)	4196 (61.3)		1819 (26.6)
Q16c …the expected results?	92 (1.4)	203 (3.0)	784 (11.6)	3646 (54.1)		2009 (29.8)
Q16d …the possible risks etc.?	199 (3.09)	312 (4.6)	615 (9.1)	3386 (50.2)		2230 (33.1)
Q17 Doctor spent enough time with you	48 (0.6)	106 (1.3)	1224 (14.8)	6869 (83.3)		
	Yes	No				
Q18 Medication prescribed	2919 (37.5)	4873 (62.5)				
	Yes	No				NA
Q19 Information about how to take medicines	2450 (79.1)	190 (6.1)				459 (14.8)
Q20 Information about side-effects	1778 (58.1)	704 (23.0)				580 (18.9)
	Not at all	Not particularly	Yes, to some extent	Yes, certainly		No contact
Q21 Nurse polite and respectful	11 (0.1)	14 (0.2)	318 (3.8)	7827 (94.0)		161 (1.9)
Q22 Could ask nurse questions	39 (0.5)	110 (1.3)	639 (7.7)	7163 (86.7)		312 (3.8)
	Not at all	Not particularly	Yes, to some extent	Yes, certainly	No contact	No need
Q23 Nurse explained procedures	66 (0.8)	147 (1.8)	530 (6.4)	4962 (60.0)	901 (10.9)	1663 (20.1)
	Yes	No	Don't know			
Q24 Information on patients' rights available	3952 (49.6)	610 (7.7)	3406 (42.7)			
	Yes	No				
Q25 Opinion/results given immediately	5289 (65.2)	2820 (34.8)				
	Not at all	Not particularly	Yes, to some extent	Yes, certainly		
Q26 Clear how the treatment would proceed	100 (1.2)	180 (2.2)	1152 (14.1)	6757 (82.5)		

Legend: NA = not applicable. *Only abbreviated items are shown. Percentages are in parentheses and are based on valid responses (including "not applicable")

The subsequent psychometric analyses of the doctors' work (DW) and nurses' work (NW) scales were based on data from 8,355 participants for DW and 8,206 participants for NW.

We evaluated the psychometric quality of the items that can be combined into the composite evaluations of the DW and NW by means of factor analysis. Except for the chi-square test, the fit of the one-factor model was very good for the DW scale: $X^2(21)=61.3$, p<0.001; CFI=0.999, RMSEA=0.025 (90% CI: [0.018; 0.033]), SRMR=0.013. The fit indices are not available for the NW scale because it consists of only three items. The values of the coefficients omega and alpha were .97 and .91, respectively, for the DW scale, and .77 and .74, respectively, for the NW scale. Figure 1 shows the factor loadings with 95% confidence intervals for both scales. The ranges of standardized loadings were .75-.92 for the DW scale and .83-.93 for the NW scale, respectively. Therefore, all items can be considered good indicators of the patients' experience with the work of the doctors and nurses. In addition, the differences between the items were relatively small.

Because the one-factor model performed well and the factor loadings were similar in magnitude, both sets of items were subsequently analysed according to the partial credit Rasch model. For the DW scale, the model fit was satisfactory, although not perfect: items Q14 and Q17 had significant item-fit test statistics (p<.001) and slightly elevated OUTFIT values (1.24 and 1.18), although they did not exceed the 1.3 criterion value. For the NW scale, the fit of all items was good (no significant test statistics, and all INFIT/OUTFIT values below 1).



Legend: Numbered circles indicate transition points where the two adjacent categories are equally likely to be chosen.

Figure 2. Person-item maps for the doctor's work scale (above) and the nurses' work scale (below).

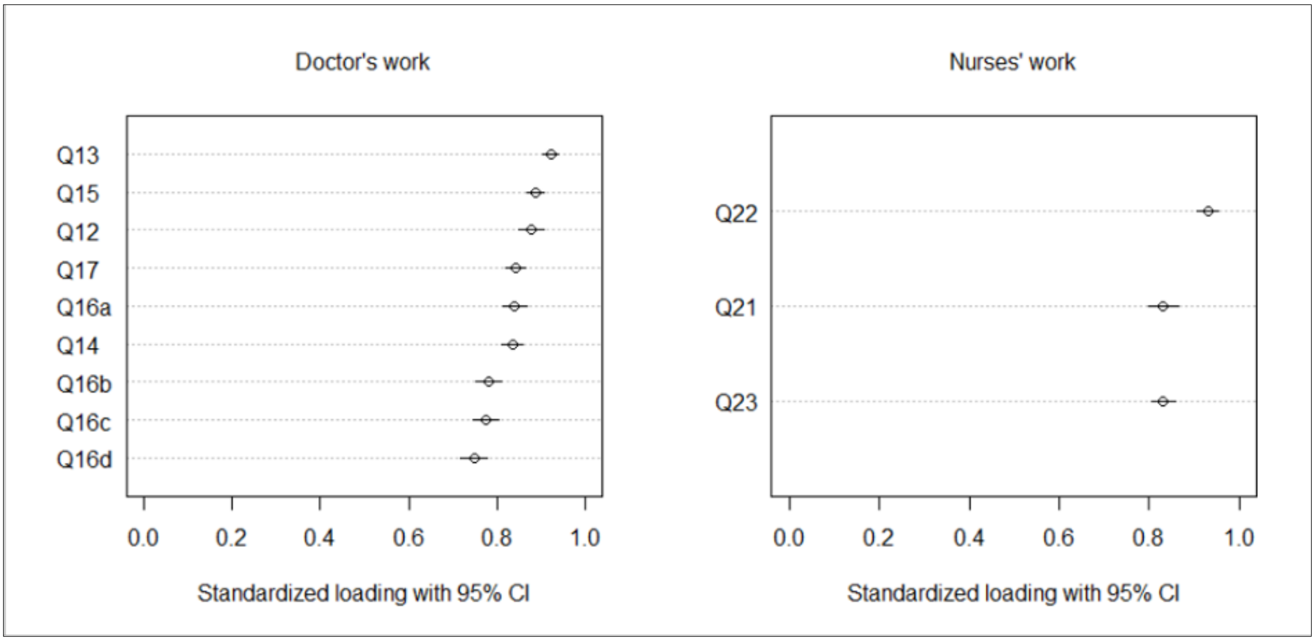


Figure 1. Standardized factor loadings (circles) with 95% confidence intervals (lines). Left: doctor's work; right: nurses' work.

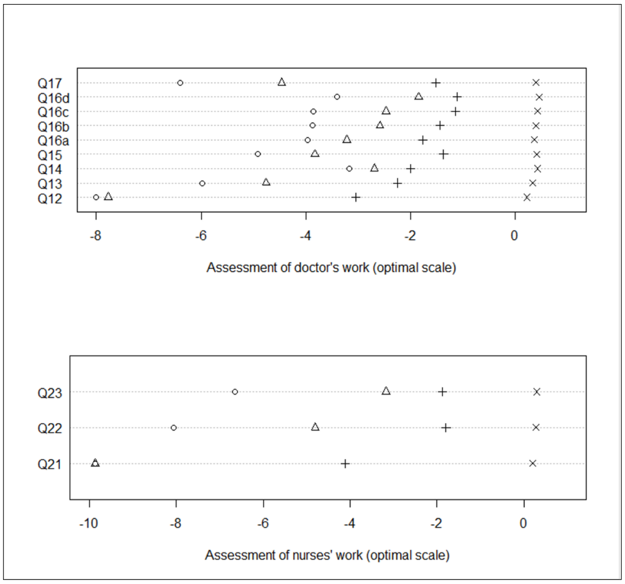
Figure 2 presents the person-item maps for the two scales. The upper part of the map shows the distribution of the scaled assessment (that is, participants' scale scores), and the lower part shows the category thresholds. The latter are the points at which the two adjacent responses are equally likely. For instance, the point labelled as 2 refers to the scale score where the person with such score will equally likely respond with "not particularly" and "yes, to some extent". For more information on person-item maps see, for instance, (18). In both cases, this distribution of person scores was highly asymmetric, reflecting a high proportion of favourable responses.

The item category thresholds mostly cover the lower part of the person distribution (especially for the DW scale), indicating that the items are more informative for persons with a less favourable experience. For items Q12-Q15, Q17, and Q21 the first two thresholds were reversed. Technically, this indicates the "ease" of the second category transition in comparison to the first one. In our case, this may also indicate a low discriminability between the lowest two categories.

To gain more insight into this issue, we analysed both sets of items with the optimal scaling using correlational aspects. Figure 3 presents the scale values of subsequent categories. While the scale values for the most favourable response ("yes, certainly") were always clearly distant from the remaining ones, the scale values for the remaining responses were sometimes very close. For item Q21, for instance, the scale values for the two lowest responses coincided, and for item Q12 they were almost the same. Thus, the lower response categories indeed appear to be empirically less clearly distinguishable compared to the highest category.

Table 2 presents the descriptive statistics for the scale sum-scores and for the general experience rating (Q27). Although the full range of possible scores was found in the data, the distribution of all three variables was very notably skewed and peaked. This was mostly due to a high proportion of high ratings. Indeed, for each variable, more than half of the participants had the highest possible value (36, 12, and 10, respectively).

To assess the validity of measures obtained with the questionnaire, we calculated the correlations between the final general assessment of treatment and the scores on the DW (τ=.54) and the NW (τ=.40) scales, respectively. The correlation between both scale scores was .43. We calculated Kendall correlations (τ) because all distributions were highly skewed and the relationships were nonlinear. All correlations were statistically significant (p<.001).



Legend: Symbols denote the scale values of subsequent categories: o = "not at all" … x = "yes, certainly".

Figure 3. Category scale values for the optimally scaled item responses.

4 DISCUSSION

The aim of this study was to investigate whether the Slovenian outpatient PREM instrument adequately captures patient reported experiences of healthcare.

We analysed in more detail the two sets of items where it made sense to create a composite score, namely DW and NW. The DW and NW scale scores generally possess good psychometric properties in terms of fit to psychometric models, magnitude of factor loadings, and reliability. Because the NW consists of only three items, the reliability of its score is somewhat lower, but still acceptable for group-level analyses. With regard to the

Table 2. Frequencies and percentages of item responses.

Item*	*n*	*M*	M_{trim}	*Mdn*	*SD*	Min	Max	Skewness	Kurtosis
doctor's work	3002	33.88	34.85	36	3.93	9	36	-2.71	8.33
nurses' work	5628	11.64	11.91	12	0.98	3	12	-3.72	17.25
general rating	8307	9.35	9.62	10	1.23	0	10	-3.04	12.65

Legend: The scale scores were calculated only for persons without NA responses. M_{trim}=10% trimmed mean.

accuracy of the measurement it should be borne in mind that the Rasch analysis implies that both scales are more informative for patients with unfavourable experience than for those with favourable experience. In our view, this is not a serious limitation, because the negative appraisals point to the aspects of treatment that should be improved. The composite scale scores are especially useful when a reliable general assessment of the doctor's or nurses' work is desired. For a more detailed evaluation of specific aspects, the individual item responses may still be of interest.

Although we did not have any data on external criterion variables at our disposal, the results of the factor and Rasch analyses, and the meaningful correlations between the single-item general assessment of treatment on one hand and the DW and NW scale scores on the other, support the validity of the questionnaire. We used the factor analysis to understand the internal structure of a set of items, and the extent to which the relationships between the items are internally consistent (26). With regard to the factor loadings, some authors suggest that loadings below .30 could be considered low (implying such items are inadequate measures of the construct), whereas the items with loadings above .30 would usually be considered as useful (27). In our case, both DW and NW items had factor loadings well above .70.

Reversed Rasch thresholds, similar category scale values, and low frequencies of unfavourable response categories were found for several items. This suggests that a response scale could be simplified to a three-point scale, especially when concise reporting of the results is desired. However, as long as a high discrimination of unfavourable experiences is of interest, the existing response options can be used.

The results are consistent with existing evidence on experience questionnaires. First, the fact that many items are highly negatively skewed is also observed in other experience questionnaires (28-30), and has some face validity as things much more often go right than they go wrong. Second, the scales DW and NW consist of items on the interaction between patient and professional, i.e., communication and exchange of information. Such topics typically lend themselves to scale development in patient experience surveys (29, 30), so the fact that these scales were also valid and reliable for the present questionnaire resonates with previous work on patient experience surveys.

4.1 Strengths and limitations

A major strength of our study is the large sample size, which implies more stable statistical estimates, including more accurate estimates of reliability coefficients, factor loadings and other model parameters. Since the instrument was developed in Slovenia, rather than being a translation of a foreign instrument, it may be assumed to be better adapted to the needs of the patients in the Slovenian healthcare system.

However, some limitations should be kept in mind when interpreting the results. Although all patients treated in the selected period were invited, it was not possible to control whether a nurse handed the invitation to participate in the survey to each patient, as foreseen in methodology. Since the participation was voluntary, different subgroups of patients could participate to a lesser or greater. The patients who were asked to participate the survey but chose not to might have answered differently to our questions. This means that sampling was not probabilistic in the strict sense. In the case that non-participants differed systematically from study participants, the final sample may suffer from selection bias. In the future, legal provision for random individual patient sampling from eReferral and e-SZBO (outpatient specialist visit database) is planned to avoid or at least reduce this limitation.

The relatively low response rate which might have influenced the composition of our sample is another limitation of the study. This is not uncommon in PREM studies with a similar methodology (i.e. inviting patients without additional reminders), where it is very difficult to achieve high response rates (31, 32). For example, in a Dutch study with a consumer quality index as a standardized survey method to measure patient experience of chronic dialysis care had a combined response rate that was higher, at 48%, but the methodology allowed for up to three reminders where necessary (33).

In our study, older participants and especially male participants (see online Appendix A, Table A1) were underrepresented in the sample, compared to the total sampling frame. While the size of the sociodemographic differences between the sampling frame and the actual sample is not dramatic, these deviations, in combination with a possible selection bias, limit the generalizability of the results. In particular, the average scale values and other descriptive statistics should be taken with caution; for instance, an NHS study carried out in the UK (34) reports that the underrepresented groups tend to be the ones with more negative experiences of care. On the other hand, the focus of this study was not the estimation of the (average) experiences in the population of patients, but the psychometric validation, and the psychometric indices are not affected by small or moderate shifts of average values. Response rates are reported to be only weakly associated with non-response bias in surveys that adhere to high standards of survey methodology (35), and surveys with response rates typical of those in public sector surveys (e.g., 35–40%) are often regarded as acceptable for the purpose of routine healthcare monitoring (36).

Similar problems were encountered in other comparable studies. For example, a Norwegian study (37) found that non-respondents were more likely to be younger and male. In the British NHS study (34), participants in online the surveys tended to be younger and better educated than participants who respond by other survey methods. These authors thus recommended the use of different alternative completion methods to mitigate the non-response bias and achieve a representative sample. In our study, we used a combination of online and paper surveys to improve the response rate and the representativeness of the sample. In this regard it should be noted that the use of the different data collection modes may also introduce biases. The phenomenon that the data collection mode affects the responses of participants is known as the measurement bias (or the lack of the measurement invariance) in the psychometric literature (38). Measurement bias can be caused by differences in design, perceived social desirability of responses and other irrelevant factors. This is a serious problem because it can undermine the comparability of responses of persons responding participating in different data collection modes. In our case, we do not see any reasons to expect a notable measurement bias: the responding in both data collection modes was anonymous and performed at home. The graphical designs of the online and paper surveys were also made as similar as possible. The previous PREM studies that were available to us did not report measurement bias in relation to the paper vs. online survey methods. Nevertheless, we compared the distributions of scale scores and the correlational structure in both sub-samples and found that they were practically the same, supporting the assumption of the measurement invariance. For details see the online supplement (https://osf.io/4guch/?view_only=1c3474c36e95481b959dd22a2cd03460).

The “not applicable” responses often cannot be avoided. The composite scores cannot be meaningfully calculated for patients who gave NA responses; in such cases, the responses should be interpreted at the item level. For research and evaluation purposes imputation techniques may be used. In such cases it should be remembered that the imputed values are only approximations used to obtain an estimate of the general assessment. In performing the analyses, we compared the results with respect to various treatments of missing data (pairwise estimation, listwise deletion, and imputation using the random forest technique) and found no notable differences.

There is a large variability in both the number and type of validity and reliability testing undertaken for the PREM instruments (6). We opted for factor analysis and the Rasch model. The results of the factor structure and the reliability of the DW and NW scales, and, in a more indirect manner, the distributions of the responses on the evaluation items not grouped into scales, suggest that the instrument is able to adequately capture patient reported experiences in intended clinical settings and populations. In conclusion, our study is an important step forward in the implementation of a long-term system for monitoring patient experience. Patient sampling will be improved in the future by using eReferral and e-SZBO databases as sampling frames. This would reduce the burden on healthcare providers (e.g., delivering envelopes), while giving us better control over the distribution of the questionnaires, increasing the quality of the data obtained. The response rate may be additionally improved by establishing an online platform, where the data collected through PREM surveys would be published and thus be useful to patients.

CONFLICTS OF INTEREST

The authors declare that no conflicts of interest exist.

FUNDING

The study was financed by the Ministry of Health of the Republic of Slovenia.

ETHICAL APPROVAL

Ethical approval for the study was obtained from the National Medical Ethics Committee of the Republic of Slovenia (NMEC), No. 0120-47/2021/4.

AVAILABILITY OF DATA AND MATERIALS

All data and materials used in this study were collected from publicly available sources and are available upon reasonable request. The questionnaire is available at the following web page: https://www.nijz.si/sites/www.nijz.si/files/uploaded/spremenjena_kopija_vprasalnik_o_izkusnjah_pacientov_z_obravnavo_v_specialisticni_ambulanti_25022019_tisk.pdf.

REFERENCES

1. Goodwin N. Towards people-centered integrated care: From passive recognition to active co-production. Int J Integr Care. 2016;16(2):15. doi: 10.5334/ijic.2492.
2. Constand MK, MacDermid JC, Dal Bello-Haas V, Law M. Scoping review of patient-centered care approaches in healthcare. BMC Health Serv Res. 2014;14:271. doi: 10.1186/1472-6963-14-271.
3. Carinci F, Van Gool K, Mainz J, Veillard J, Pichora EC, Januel JM, et al. Towards actionable international comparisons of health system performance: expert revision of the OECD framework and quality indicators. Int J Qual Health Care. 2015;27(2):137-146. doi: 10.1093/intqhc/mzv004.

4. De Rosis S, Cerasuolo D, Nuti S. Using patient reported measures to drive change in healthcare: The experience of the digital, continuous and systematic PREMs observatory in Italy. BMC Health Serv Res. 2020;20:315. doi: 10.1186/s12913-020-05099-4.
5. Bull C, Byrnes J, Hettiarachchi R, Downes M. A systematic review of the validity and reliability of patient reported experience measures. Health Serv Res. 2019;54(5):1023-1035. doi: 10.1111/1475-6773.13187.
6. Organization for Economic Co-operation and Development. Recommendations to OECD ministers of health from the high-level reflection group on the future of health statistics [Internet]. 2017 [cited 2022 Mar 15]. Available from: https://www.oecd.org/health/Recommendations-from-high-level-reflection-group-on-the-future-of-health-statistics.pdf
7. Fernandes S, Fond G, Zendjidjian X, Michel P, Baumstarck K, Lancon C, et al. The Patient reported experience measure for improving quality of care in mental health (PREMIUM) project in France: Study protocol for the development and implementation strategy. Patient Prefer Adherence. 2019;13:165-177. doi: 10.2147/PPA.S172100.
8. Fujisawa R, Klazinga N. Measuring patient experiences (PREMs): Progress made by the OECD and its member countries between 2006 and 2016. OECD Health Working Papers, No. 102. OECD Publishing: Paris; 2017. doi: 10.1787/893a07d2-en.
9. Sjetne, I, Bjertnaes O, Olsen RV, Iversen HH. The Generic Short Patient Experiences Questionnaire (GS-PEQ): Identification of core items from a survey in Norway. BMC Health Serv Res. 2011(1):88. doi: 10.1186/1472-6963-11-88.
10. OPEQ, Garratt AM, Bjaertnes OA, Krogstad U, Gulbrandsen P. The OutPatient Experiences Questionnaire (OPEQ): Data quality, reliability, and validity in patients attending 52 Norwegian hospitals. Qual Saf Health Care. 2005;14:433-437. doi: 10.1136/qshc.2005.014423.
11. CAHPS, The American Consumer Assessment of Healthcare Providers and Systems (CAHPS) [Internet]. 2012 [cited 2023 April 3]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3748811/pdf/nihms391989.pdf
12. Delnoij DM, Rademakers JJ, Groenewegen PP. The Dutch Consumer Quality Index: an example of stakeholder involvement in indicator development. BMC Health Serv Res. 2010;10:88. doi: 10.1186/1472-6963-10-88.
13. OECD. Health at a glance [Internet]. 2021 [cited 2023 April 4]. Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2021_ae3016b9-en
14. Ministrstvo za zdravje. Nacionalna anketa [Internet]. 2007 [cited 2023 April 3]. Available from: https://www.pb-begunje.si/gradiva/Rezultati2007135143657636.pdf
15. Murko E, Kralj M, Poldrugovac M, Ropret N, Šetinc M, Zaletel M. Izkušnje pacientov z zunajbolnišnično specialistično zdravstveno obravnavo: raziskava v Sloveniji med odraslimi pacienti. Javno zdravje. 2021;1:1-9. doi: 10.26318/JZ-2021-1.
16. Pettersen K, Veenstra M, Guldvog B; Kolstad A. The Patient Experiences Questionnaire: Development, validity and reliability. Int J Quality Health Care. 2004;16(6):453-463. doi: 10.1093/intqhc/mzh074.
17. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Modeling. 1999;6(1):1–55. doi: 10.1080/10705519909540118.
18. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences. 2nd ed. Mahwah: Lawrence Erlbaum Associates Publishers; 2007.
19. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2021.
20. Rosseel Y. lavaan: An R package for structural equation modeling. J Stat Softw. 2012;48(2):1-36. doi: 10.18637/jss.v048.i02.
21. Kelley K. MBESS: The MBESS R Package. R package version 4.9.0 [Internet]. 2022 [cited 2023 April 3]. Available from: https://CRAN.R-project.org/package=MBESS
22. Mair P, Hatzinger, R, Maier MJ. eRm: Extended rasch modeling 1.0-2 [Internet]. 2021 [cited 2023 April 3]. Available from: https://cran.r-project.org/package=eRm. 2021
23. Mair P, De Leeuw J. aspect: A general framework for multivariate analysis with optimal scaling. R package version 1.0-5. [Internet]. 2018 [cited 2023 April 3]. Available from: https://cran.r-project.org/package=aspect
24. Revelle W. psych: Procedures for personality and psychological research, version 2.1.9. [Internet]. 2021 [cited 2023 April 3]. Evanston: Northwestern University; 2021. Available from: https://CRAN.R-project.org/package=psych
25. Stekhoven DJ. missForest: Nonparametric missing value imputation using random forest. R package version 1.5. 2022.
26. McCoach DB, Gable RK, Madura JP. Instrument development in the affective domain: School and corporate applications. 3rd ed. New York: Springer; 2013.
27. Raykov T, Marcoulides GA. Introduction to psychometric theory. New York: Routledge, Taylor & Francis; 2011.
28. Bruster S. The Picker Patient Experience Questionnaire: Development and validation using data from in-patient surveys in five countries. Int J Qual Health Care. 2002;14(5):353–358. doi: 10.1093/intqhc/14.5.353.
29. de Boer D, Delnoij D, Rademakers J. The discriminative power of patient experience surveys. BMC Health Serv Res. 2011;11:332. doi: 10.1186/1472-6963-11-332.
30. Hays RD, Skootsky SA. Patient experience with in-person and telehealth visits before and during the COVID-19 pandemic at a large integrated health system in the United States. J Gen Intern Med. 2022;37(4):847-852. doi: 10.1007/s11606-021-07196-4.
31. Damman OC, Hendriks M, Rademakers J, Spreeuwenberg P, Delnoij DM, Groenewegen PP. Consumers' interpretation and use of comparative information on the quality of health care: the effect of presentation approaches. Health Expect. 2012;15(2):197-211. doi: 10.1111/j.1369-7625.2011.00671.
32. OECD. Measuring what matters: The patient reported indicator surveys: Patient reported indicators for assessing health system performance [Internet]. 2019 [cited 2023 April 17]. Available from: https://www.oecd.org/health/health-systems/Measuring-what-matters-the-Patient-Reported-Indicator-Surveys.pdf
33. Van der Veer SN, Jager KJ, Visserman E, Beekman RJ, Boeschoten EW, de Keizer NF, et al. Development and validation of the Consumer Quality index instrument to measure the experience and priority of chronic dialysis patients. Nephrol Dial Transplant. 2012;27(8):3284-3291. doi: 10.1093/ndt/gfs023.
34. NHS, Care Quality Commission. Adult inpatient survey: quality and methodology report [Internet]. 2021 [cited 2023 April 17]. Available from: https://www.cqc.org.uk/sites/default/files/2022-09/20220920_aip22_QualityMethodology.odt
35. Johnson TP, Wislar JS. Response rates and nonresponse errors in surveys. JAMA. 2012;307(17):1805–1806. doi: 10.1001/jama.2012.3532.
36. Faraz A, Burt J, Roland M. Measuring patient experience: Concepts and methods. Patient. 2014;7:235–241. doi: 10.1007/s40271-014-0060-5.
37. Garratt AM, Bjørngård JH, Dahle KA, et al. The Psychiatric Out-Patient Experiences Questionnaire (POPEQ): Development, reliability and validity. Nordic J Psychiatry (in press).
38. Millsap RE. Statistical approaches to measurement invariance. New York: Routledge; 2011.