

Method

Cite this article: Poder TG, Coulibaly LP, Hassan Al, Conombo B, Laberge M (2022). Test–retest reliability of the Cost for Patients Questionnaire. *International Journal of Technology Assessment in Health Care*, **38**(1), e65, 1–8
<https://doi.org/10.1017/S0266462322000460>

Received: 18 November 2021

Revised: 27 May 2022

Accepted: 20 June 2022

Key words:

Test–retest; Reliability; Cost for patients questionnaire; Intraclass correlation coefficient; Cohen's Kappa coefficient





Author for correspondence:

*Maude Laberge,

E-mail: maude.laberge@fsa.ulaval.ca

The authors would like to thank the Quebec Support Unit of the Strategy for Patient Oriented Research (SPOR) for their financial support to conduct this study. T.G.P. is a member of the FRQS-funded Centre de Recherche de l'USMM. M.L. is a member of the Centre de Recherche du CHU de Québec, and of VITAM, Centre de Recherche en Santé Durable, both funded by the FRQS. T.G.P. and M.L. are fellows of the FRQS.

Test–retest reliability of the Cost for Patients Questionnaire

Thomas G. Poder^{1,2} , Lucien P. Coulibaly^{3,4} , Abakar Idriss Hassan⁵,
Blanchard Conombo^{6,7}  and Maude Laberge^{7,8,9*} 

¹Département de Gestion, Évaluation et Politique de Santé, École de Santé Publique de l'Université de Montréal, Montréal, QC, Canada; ²Centre de Recherche de l'Institut Universitaire en Santé Mentale de Montréal, Montréal, QC, Canada;

³Faculté des Lettres et Sciences Humaines, Université de Sherbrooke, Sherbrooke, QC, Canada; ⁴Centre de Recherche sur le Vieillessement, Sherbrooke, QC, Canada; ⁵Institut National de Santé Publique, Québec City, QC, Canada; ⁶Department of Social and Preventative Medicine, Université Laval, Québec City, QC, Canada; ⁷Population Health and Optimal Health Practices Research Unit, Centre de Recherche du CHU de Québec, Université Laval, Québec City, QC, Canada;

⁸Department of Operations and Decision Systems, Faculty of Administration, Université Laval, Québec City, QC, Canada and ⁹VITAM—Centre de Recherche en Santé Durable, Université Laval, Québec City, QC, Canada

Abstract

Objectives: To investigate the test–retest reliability of the Costs for Patients Questionnaire (CoPaQ).

Methods: Through an online survey, individuals were invited to participate in a two-step study to assess the test–retest reliability of the CoPaQ. Participants to the first step were invited to complete the questionnaire a second time 2 weeks after. Reliability was assessed by calculating Cohen's Kappa coefficients and intraclass correlation coefficients (ICC) for discrete and continuous data, respectively. A sensitivity analysis was carried out.

Results: From a total of 1,200 participants who completed the first test, 403 completed the second test. The ICC varied from –0.00 to 0.98 with poor, moderate, good, and excellent results. The Kappa coefficients varied from –0.004 to 0.65 and were poor, slight, fair, moderate, and substantial. The sensitivity analysis showed the median value of ICC and Kappa coefficients for each category before and after the outliers' exclusion. The median value of ICC changed from 0.30 (before) to 0.70 (after), and from 0.12 (before) to 0.04 (after), respectively, for each category. The median value of the Cohen's Kappa coefficient increased from 0.44 (before) to 0.46 (after) and decreased from 0.32 (before) to 0.30 (after), respectively.

Conclusions: Test–retest reliability results indicated that the CoPaQ has a moderate reliability in terms of ICC and Kappa coefficients. The moderate reliability observed gives additional support for the applicability of this tool in economic evaluations of health interventions. Additional studies including on other properties and a cultural adaptation could further enhance the use of the tool.

In recent years, there has been growing concern about integrating the patient perspective in research, especially in the decision-making and study design processes including in the choice and assessment of clinical and health outcomes (1;2). There is also a need to consider the patient's perspective in economic evaluations within healthcare systems. Patient questionnaires are often used to collect data about their utilization of healthcare services and the costs of that utilization from the perspective of a healthcare organization or public and private insurers. However, there is seldom consideration and collection of costs from the patient and the societal perspective, despite recognition that costs of health care services for the patients may affect their utilization and their outcomes (3). Data of good quality are essential for the development of appropriate health policies and interventions (4). Data quality refers to a measure of the condition of data based on factors such as accuracy, completeness, consistency, reliability, and timeliness (5;6). Cost data are critical when conducting health economic assessments. As such, the reliability of cost data collection methods is particularly important in the context of public policy supporting healthcare management, and with the purpose of building public awareness about the factors that affect health (6;7). The use of patient questionnaires confirms that patients' concerns are increasingly considered in decision making. When considering the patient perspective in economic evaluations, costs incurred by patients need to be measured with appropriate and adapted measurement instruments. As with many questionnaires, cost instruments should undergo psychometric testing and the results should meet accepted standards of reliability, validity, and responsiveness prior to their use in any assessment (8). Without these properties, instruments will unlikely be widely used in research studies (9;10). The reliability plays a central role in developing interpretative validity evidence in general and for the estimation of validity coefficients specifically (11).

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Reliability refers to (i) the ability of a scale to yield reproducible and consistent results and (ii) the extent to which a questionnaire's score is free of random error (12;13). An essential element of reliability is that the scores on tests are consistent when they are obtained under similar testing conditions (11). The two relevant forms of reliability assessment are internal consistency (where "repeated measurement" is conceptualized as multiple "replicates" or items in a single administration) and test–retest reliability (which concerns consistency of scores across two separate measurements over time and is sometimes referred to as stability or reproducibility) (14). In health research, test–retest reliability is of greater recommended reliability method because it can be used to assess temporal fluctuations (14). The recommended test–retest reliability coefficients are the intraclass correlation coefficient (ICC) for continuous variables and Cohen's Kappa coefficient for categorical variables (15).

Measurement instruments are often tested for validity and particularly content validity but another dimension that is important to measure is the reliability of an instrument. To our knowledge, there are currently very few costs questionnaires or health resource utilization instruments (16) available to measure the different types of health-related costs from the patient's perspective that reported on the test–retest reliability in the context of economic evaluations. Recently, a comprehensive tool, namely the "Costs for Patients Questionnaire" (CoPaQ), was developed by Laberge *et al.* (17) to measure out-of-pocket costs for patients and their caregivers. The CoPaQ was intended to be applied to a diverse population of patients and is not condition specific. The development of the CoPaQ was conducted with the objective that it would be generalizable to ambulatory care patients in different healthcare systems. This tool measures the nonreimbursed costs (i.e., direct medical, direct nonmedical, and indirect) of a health condition for patients and their informal caregiver (17). The CoPaQ is thus recommended to researchers who wish to capture this category of costs in a standardized way. Before this measurement instrument can be used for research or economic evaluations, its reliability must be established (9;18). The purpose of this study was to investigate the test–retest reliability of the CoPaQ.

Methods

Study Design and Participants

A test–retest reliability study design was conducted. To participate in the study, subjects had to be an adult (aged 18 years or older) and a patient meeting the following criteria: (i) have used health services during the reference period (i.e., between 1 Nov 2019 and 31 Jan 2020); (ii) have a condition that requires using healthcare services; and (iii) live in Quebec, Canada. Patients living in institutions were excluded from this study. The survey was conducted in French, the original language in which the CoPaQ was developed.

Sample Size Calculation

A small pilot test–retest was carried out in a previous study with eighteen participants. Based on these pilot data (17) and the number of people in Quebec, 385 participants were targeted in the recruitment. This sample size was estimated by setting respectively the statistical power and alpha risk at 95 and 5 percent for a Quebec population of 6.9 million adults. To achieve the target sample size and reduce potential missing data, 400 participants were targeted. We also oversampled to three times this sample size (400×3) for the first part of the test since the survey company indicated that

there could be a large loss of participants between the test and the retest; generally, between 30 and 40 percent of their panelists participate in a second survey (informal communication). Finally, a total of 1,200 participants were invited to complete the questionnaire at the test. The same participants were then asked to complete the questionnaire again 2 weeks later, that is, at the retest. We closed the survey once the target sample size of 400 was reached.

Recruitment and Data Collection

Respondents were recruited among the online panel members of the survey company Dynata (Plano, TX) and were structured to achieve a random sample according to our eligibility criteria. Participants do not receive a compensation for their participation, but they earn points from Dynata's partners redeemable for discounts, special offers, and member-only promotions. Data were collected from 14 Feb 2020 through 3 Mar 2020 via an online questionnaire. To carry out the CoPaQ's test–retest reliability, the same respondents were invited to self-complete the electronic CoPaQ on two occasions, 2 weeks apart, but covering the same period for their costs. A 2-week interval was thought to be long enough for respondents do not simply remember their previous response, but short enough to avoid any major change in condition that could affect their perception of costs and to reduce memory bias.

Instrument/Questionnaire

The CoPaQ is a comprehensive tool to measure direct and indirect out-of-pocket costs of a health condition for patients and their families to various outpatient contexts (17). This tool was developed after a systematic review (19). The objective of this systematic review was to identify validated tools to measure costs associated with a health condition from a patient's perspective as well as the different components of health-related costs. Nine partially validated tools targeting distinct patient populations were found (19). These tools measure costs for patients (direct and indirect), intangible costs, and costs for caregivers. Among these studies, few adopted a rigorous development and validation process, as evidenced by the results of our review (19). In conducting the systematic review, we extracted all the cost elements used in the identified studies to make a preliminary list for the Delphi. The Delphi process involved fourteen panelists: six researchers with a clinical or health economics expertise and eight patients (17). Following the Delphi process, a small pilot test–retest was carried out with eighteen new participants on two different occasions separated by a 2-week interval (17). The participants of the pilot met the same eligibility criteria as patients from the Delphi panel (17). The pilot test–retest allowed to calculate the ICC and Kappa coefficients and to identify the items that lack clarity or that may not be appropriate for, or discriminate between respondents (10). Six items that presented comprehension problems were identified and subsequently revised, and three items were removed (17). The CoPaQ is composed of forty-one items and subdivided into eight categories: (i) the patient's costs (sixteen items), (ii) the time spent on accessing medical services (two items), (iii) the costs related to workforce participation (two items), (iv) the patient's financial distress (one item), (v) the informal caregiver's costs (seven items), (vi) the informal caregiver's timer not directly related to medical services (three items), (vii) the patient's sociodemographic characteristics (nine items), and (viii) a note to researchers (web link to complete the EQ-5D-5L questionnaire, one item). We calculated ICC (Table 1) and Cohen's Kappa coefficients (Table 2) on eight and

Table 1. Intraclass Correlation Coefficient (ICC) Results

Items	ICC (95% CI)	p-value	ICC interpretation
1.3. On average, how many kilometers (round trip) did you travel to get to the health center or for your consultations?	0.76 (.71; .80)	<.0001	Good
1.5. On average, how long did you wait in the clinic?	0.70 (.63; .74)	<.0001	Moderate
2.1. How much time did you spend traveling to and from the health center or for your consultations (round trip)?	0.21 (.11; .31)	<.0001	Poor
2.2. Approximately how long did you need to wait to receive medical services (e.g., over the phone, or to schedule an appointment at the clinic prior to your consultation)?	0.30 (.19; .38)	<.0001	Poor
3.2. What is your rough estimate (net amount) of the incurred loss of income?	0.98 (.96; .99)	<.0001	Excellent
6.1. How much do you estimate the total time spent on travel (round trip) by your informal caregiver or companion to accompany you to the healthcare center?	0.04 (−.46; .56)	.435	Poor
6.2. What is the estimated average time per week your caregiver or the person accompanying you spends performing various tasks (e.g., housework, home care)?	0.76 (.48; .90)	<.0001	Good
6.3. How long do you estimate the waiting time for your informal caregiver or companion during all your medical consultations?	−.00 (−.38; .41)	.506	Poor

Table 2. Kappa Results

Items	Agreement (%)	Expected agreement (%)	Kappa value	Standard error (SE)	Pr > Z	Kappa interpretation
1.1. Did you travel to a health center (e.g., hospital, family medicine group, physiotherapy clinic) to receive healthcare services, or for consultations?	91.60	89.53	0.19	0.14	<.0001	Slight
1.2. What means of transportation did you use to get to the health center or to your consultations?						
1.2.1. Public transit (bus, metro/subway)	88.91	71.23	0.61	0.06	<.0001	Substantial
1.2.2. Taxi	89.71	82.82	0.09	0.40	<.0001	Slight
1.2.3. Your personal vehicle	84.70	59.54	0.62	0.04	<.0001	Substantial
1.2.4. Other means of transportation (on foot, by bicycle, personal vehicle of the person who went with you)	87.07	72.93	0.52	0.06	<.0001	Fair
1.4. Did you ever pay for parking during your visits?	83.04	52.15	0.65	0.04	<.0001	Substantial
1.6. When traveling to the health center or to consultations, did you pay for accommodation?	88.91	86.10	0.20	0.11	<.0001	Slight
1.7. Did you ever pay any portion “out of pocket” for your prescribed medication that was not reimbursed?	76.51	56.62	0.46	0.05	<.0001	Moderate
1.8. Did you pay for nonprescribed medication or dietary supplements (e.g., aspirin, natural products)?	73.61	50.78	0.46	0.05	<.0001	Moderate
1.9. Did you incur any expenses for home care services (e.g., rehabilitation, etc.)?	97.62	97.13	0.17	0.27	.000159	Slight
1.10. Did you incur expenses for the purchase of any medical devices (e.g., blood pressure monitor, blood glucose monitor, walker, wheelchair, raised toilet seat, protective underwear, shower rails)?	90.80	80.50	0.53	0.08	<.0001	Fair
1.11. Did you renovate your home in order to better accommodate your condition?	98.41	98.42	−0.004	0.40	.908	Poor
1.12. Did you pay for any additional nonmedical services during or following your consultations (e.g., insurance forms, sending photocopies, doctor's certificate)?	82.90	81.77	0.61	0.08	<.0001	Substantial
1.13. Did you pay for any additional nonmedical services during or following your consultations (e.g., insurance forms, sending photocopies, doctor's certificate)?	80.50	67.13	0.40	0.06	<.0001	Fair
1.14. Did you pay for any nonmedical care services (e.g., physiotherapy, occupational therapy, psychology, osteopathic treatments, massage therapy, dentistry, or optometry)?	80.57	62.36	0.48	0.05	<.0001	Moderate

(Continued)

Table 2. (Continued)

Items	Agreement (%)	Expected agreement (%)	Kappa value	Standard error (SE)	Pr > Z	Kappa interpretation
1.15. Did you pay for someone to care for your dependents during any of your consultations (e.g., childcare or pet care)?	98.83	98.25	0.33	0.33	<.0001	Fair
1.16. Did you incur any other expenses (e.g., food services, any specific meals related to accessing healthcare services)?	90.50	86.74	0.28	0.11	<.0001	Fair
3.1. Have you suffered a loss of income?	90.00	79.88	0.50	0.08	<.0001	Moderate
4.1. I feel financially stressed due to my state of health	61.47	30.30	0.45	0.04	<.0001	Moderate
5.1. Did a caregiver or anyone else accompany you to your consultations at the health center?	84.79	77.08	0.34	0.08	<.0001	Fair
5.2. Did you travel together to the health center?	85.00	78.37	0.30	0.08	<.0001	Fair
5.5. Did the caregiver or the person accompanying you pay for any accommodations while accompanying you to the health center or to your consultations?	86.01	79.87	0.30	0.08	<.0001	Fair
5.6. Did your caregiver or the person accompanying you receive any training in order to assist you?	70.00	58.50	0.13	0.28	.07	Slight
5.7. Did your caregiver or the person accompanying you incur any other expenses while accompanying you?	73.68	55.12	0.41	0.23	.0039	Moderate

twenty-four items, respectively. We also calculated the median value of ICCs and Kappa coefficients to represent central tendencies that are not influenced by extreme scores or outliers. In total, three (6.1, 6.2, and 6.3) of the eight ICC items and three (5.5, 5.6, and 5.7) of the twenty-four Kappa items were specific to patients with an informal caregiver.

Statistical Analysis

Once we completed data collection for the test–retest, we analyzed the distribution of responses for each item to find out the potential outliers. The descriptive statistics for the sociodemographic variables (i.e., age, gender, matrimonial status, education, income, urban or rural area) were determined. A reliable (or consistent) response to a question was defined as one where the same response level was provided at both time periods (20). The reliability was assessed by calculating the Kappa coefficient and ICC. The ICC statistic is the most suitable and most commonly used reliability parameter for continuous measures (21). The ICCs are reported with their 95 percent confidence intervals and the formula is

$$ICC = [MS_P - MS_E] / [MS_P + (k - 1)MS_E + (k/n)(MS_0 - MS_E)]$$
 where MS_P is the mean square for participants, MS_E is the mean square for error, MS_0 is the mean square for observers, n is the number of participants, and k is the number of observers/measurements (18). The Kappa statistic is a measure of “strength” agreement for categorical variables. It indicates the proportion of agreement beyond that expected by chance, that is, the achieved beyond-chance agreement as a proportion of the possible beyond-chance agreement (22). The Cohen’s Kappa coefficient (k) is calculated with the following formula: $k = [Pr(a) - Pr(e)] / [1 - Pr(e)]$, where $Pr(a)$ represents the actual observed agreement, and $Pr(e)$ represents agreement by chance (12). Interpretations of ICC coefficients and Kappa values are suggested in Table 3 (18;23;24). To observe the outliers’ impact on the results, a sensitivity analysis was carried out. The analysis consisted of subdividing all the items into four categories as follows: category 1 (items

Table 3. Interpretation of Intraclass Correlation Coefficient (ICC) and Cohen’s Kappa

ICC coefficients	
Value	Interpretation
ICC < 0.5	Poor
0.5 ≤ ICC < 0.75	Moderate
0.75 ≤ ICC < 0.9	Good
0.9 ≤ ICC	Excellent
Kappa statistics	
Kappa < 0.00	Poor
0.00 < Kappa < 0.20	Slight
0.20 ≤ Kappa < 0.40	Fair
0.40 ≤ Kappa < 0.60	Moderate
0.60 ≤ Kappa < 0.80	Substantial
0.80 ≤ Kappa	Almost perfect

Koo and Li (18); Landis and Koch (23); Feinstein and Cicchetti (24).

1.3, 1.5, 2.1, 2.2, and 3.2), category 2 (items 6.1, 6.2, and 6.3), category 3 (items 1.1, 1.2 [1.2.1, 1.2.2, 1.2.3, 1.2.4], 1.4, 1.7, 1.8, 1.12, 1.14, 1.15, 1.16, 3.1, and 4.1), and category 4 (items 5.1, 5.2, 5.5, 5.6, and 5.7). We considered these subgroups because participants’ responses over time could be affected by these subgroups’ items. The categories 1 and 3 were related to items specific to the patients, and the categories 2 and 4 were related to items specific to the informal caregivers. Next, the median values of the ICC (categories 1 and 2) and Kappa coefficients (categories 3 and 4) were determined before and after the outliers’ exclusion. All analyses were performed using the statistical software R version 3.4.4 (2018-03-15). The statistical significance level was set at 5 percent for a two-tailed test.

Table 4. Characteristics of Test–Retest Participants

Variable		Test N (%)	Retest N (%)
Sex	Women	196 (51.71)	196 (51.71)
	Men	183 (48.29)	183 (48.29)
Employment	Yes	186 (49.07)	190 (50.13)
	No	193 (50.93)	189 (49.87)
Marital status	Married	127 (33.51)	128 (33.77)
	In a relationship	113 (29.81)	110 (29.02)
	Single	84 (22.16)	84 (22.16)
	Separated	4 (1.05)	4 (1.05)
	Divorced	35 (9.23)	37 (9.76)
	Widowed	16 (4.22)	16 (4.22)
Age group	18–24	14 (3.70)	15 (3.95)
	25–34	28 (7.39)	26 (6.86)
	35–44	61 (16.10)	63 (16.62)
	45–54	47 (12.40)	45 (11.87)
	55–64	105 (27.70)	107 (28.23)
	65–74	105 (27.70)	103 (27.18)
	75–84	16 (4.22)	18 (4.75)
	85+	3 (0.79)	2 (0.52)
Background	Elementary	3 (0.80)	4 (1.05)
	High school	87 (22.95)	87 (22.95)
	DEP	42 (11.08)	45 (11.87)
	College	36 (9.50)	31 (8.17)
	CEGEP	78 (20.58)	82 (21.63)
	University certificate	40 (10.55)	32 (8.44)
	Bachelor's degree	66 (17.41)	72 (18.99)
	Master's degree	24 (6.33)	23 (6.06)
Doctorate (MD, PhD)	3 (0.80)	3 (0.80)	
Gross annual income	<\$5,000	3 (0.80)	3 (0.80)
	\$5,000–9,999	13 (3.43)	10 (2.63)
	\$10,000–14,999	10 (2.63)	14 (3.70)
	\$15,000–19,999	29 (7.65)	31 (8.18)
	\$20,000–24,999	18 (4.75)	19 (5.01)
	\$25,000–29,999	15 (3.95)	16 (4.22)
	\$30,000–34,999	20 (5.27)	14 (3.70)
	\$35,000–39,999	19 (5.01)	20 (5.27)
	\$40,000–44,999	22 (5.80)	21 (5.54)
	\$45,000–49,999	31 (8.18)	25 (6.59)
	\$50,000–59,999	21 (5.54)	24 (6.33)
	\$60,000–69,999	27 (7.12)	30 (7.91)
	\$70,000–79,999	3 (0.79)	3 (0.79)
	\$80,000–89,999	31 (8.17)	27 (7.12)

(Continued)

Table 4. (Continued)

Variable		Test N (%)	Retest N (%)
	\$90,000–99,999	36 (9.50)	40 (10.55)
	\$100,000–124,999	14 (3.69)	16 (4.22)
	\$125,000–149,999	18 (4.74)	20 (5.27)
	≥\$150,000	25 (6.59)	25 (6.59)
	Missing values	24 (6.33)	21 (5.54)
Geographical area	Urban area	263 (69.40)	264 (69.66)
	Rural area	116 (30.60)	115 (30.34)

Ethics

Ethics approval was obtained from the research ethics committee of the CIUSSS de l'Estrie-CHUS (Project #2019-3102-Écosanté 2). The data collected for this study were anonymous and only an identification code provided by the survey company could identify respondents. The confidentiality of the data was respected. Completing the survey online was considered as a consent to participate.

Results

One thousand and two hundred individuals participated in the first round of the test–retest reliability, and out of these, 403 participants completed both the test and the retest with a mean completion time of 8 min (SD of 5 min). After the two rounds, twenty-four respondents (5.9 percent) were excluded from further analysis because their socio-demographic information between the test and retest was highly inconsistent. For example, when the participant reported a different sex during the two rounds. Twenty-four outlier responses were identified from six ICC items, respectively, one (item 1.3), four (item 1.5), ten (item 2.1), three (item 2.2), four (item 6.2), and two (item 6.3). Each outlier response was imputed by the average of the responses to the test or retest of each concerned item (25). Finally, the sample size used in the analysis was 379 participants, of which nineteen participants reported having an informal caregiver during their consultations at the healthcare center. The mean age of the respondents was 55 years, 48.29 percent were men and 51.71 percent were women and, 63.32 percent were married or partnered (Table 4). Half of this population had no paid job at the time of the survey, and nearly one-third lived in rural areas (Table 4). More detailed information regarding respondents' characteristics is reported in Table 4. For each item of interest, only individuals that responded to the test and retest were considered to estimate the ICC and the Kappa coefficients.

The test–retest reliability results show quite varied indices of temporal stability for both ICC and Kappa items. The results indicated that the range of values of ICC was from -0.00 to 0.98 [median (0.5) and interquartile range (Q3–Q1) (0.76–0.17)] and was poor (items 2.1, 2.2, 6.1, and 6.3), moderate (item 1.5), good (items 1.3 and 6.2), and excellent (item 3.2) (Table 1). The range of values of Kappa coefficients was from -0.004 to 0.65 [median (0.41) and interquartile range (0.51–0.26)] and was poor (item 1.11), slight (items 1.1, 1.2.2, 1.6, 1.9, and 5.6), fair (items 1.2.4, 1.10, 1.13, 1.15, 1.16, 5.1, 5.2, and 5.5), moderate (items 1.7, 1.8, 1.14, 3.1, 4.1, and 5.7), and substantial (items 1.2.1, 1.2.3, 1.4, and 1.12) (Table 2).

During this study, we asked participants for their opinion on the most appropriate period that should be covered to reduce the

Table 5. Sensitivity Analyzes Results

	Before		After	
	ICC	Kappa	ICC	Kappa
Items which were specific to patients				
Median	0.30	0.44	0.70	0.46
Mean	0.41	0.39	0.59	0.40
Items which were specific to informal caregivers				
Median	0.12	0.32	0.04	0.30
Mean	0.11	0.34	0.26	0.29

ICC, intraclass correlation coefficient.

memory loss of the small ticket items. Among the respondents, 21.59 percent favored a 1-month period, 47.89 percent a 3-month period, 14.64 percent a 6-month period, 13.15 percent a 12-month period, and 2.73 percent had no opinion.

The sensitivity analyses indicated for the first category (i.e., items 1.3, 1.5, 2.1, 2.2, and 3.2), a median value of ICC at 0.30 before the exclusion of outliers. After the exclusion, the ICC ranged from 0.21 to 0.98 (median 0.70) (Table 5). For the second category (i.e., items 6.1, 6.2, and 6.3), the median value of ICC was 0.12 before the outliers' exclusion. After the exclusion, the ICC coefficients ranged from -0.00 to 0.76 (median 0.04) (Table 5). For the third category (items 1.1, 1.2 [1.2.1, 1.2.2, 1.2.3, 1.2.4], 1.4, 1.7, 1.8, 1.12, 1.14, 1.15, 1.16, 3.1, and 4.1), the median value of the Kappa coefficient was 0.44 before the outliers' exclusion. After the exclusion, the Cohen's Kappa coefficients ranged from 0.09 to 0.65 (median 0.46). For the fourth category (items 5.1, 5.2, 5.5, 5.6, and 5.7), the median value of the Kappa coefficient was 0.32 before the outliers' exclusion. After the exclusion, the Cohen's Kappa coefficients ranged from 0.13 to 0.41 (median 0.30) (Table 5).

Discussion

In this study, the reliability of the self-administrated questionnaire for measuring the costs of a health condition for patients and their families (CoPaQ) was evaluated by the test–retest method. We found that the ICC coefficients varied from -0.00 to 0.98 (median 0.5) and the Kappa coefficients varied from -0.004 to 0.65 (median 0.45). The reliability test of the CoPaQ indicated that this instrument has a moderate reliability in terms of ICC (18) and Kappa coefficients (23;24).

We cannot compare these results with any other of similar patient costs tools because none of the tools that we identified in a systematic review had undergone such reliability testing using a test–retest (19). The questionnaire's reliability was examined by using tests of internal consistency in two included studies (26;27). Therefore, the findings of our study that were estimated in terms of ICC and Cohen's Kappa coefficients are not comparable to those produced by other studies. This would appear to confirm that our study seems to be the first to assess a test–retest reliability of a cost for patients questionnaire ICCs and Cohen's Kappa coefficients.

In this study, reliability refers to the stability of the measuring instrument: a reliable instrument will obtain the same results or almost with repeated administrations of the test (28). The results of this study provide an idea of the reliability of the CoPaQ used to measure patient costs, which is of great interest for researchers conducting economic evaluations of healthcare interventions. Our

results compare well with a previous pilot test–retest conducted on the CoPaQ with a small sample ($n = 18$). The results of this pilot showed that the ICC varied from -0.02 to 0.99 (median 0.62) and the Kappa coefficients varied from -0.11 to 1.00 (median 0.86) (17). These results suggested that the pilot version of the CoPaQ had a moderate to good reliability in terms of ICC and gives consistent results between the two measurement periods in terms of Kappa coefficients. However, these results must be interpreted with caution. The median coefficient values of ICC and Kappa dropped significantly when calculated with a larger sample size. There are a number of possible explanations of this difference. First, socio-demographic characteristics of participants in the pilot were different from those of the larger sample size. For instance, 10 percent of participants in the pilot were unemployment and 5.5 percent lived in rural areas, while in the larger sample size, half had no paid job at the time of the survey, and nearly one-third lived in rural areas. Another explanation is that respondents in the pilot test–retest may have responded more carefully than those recruited by the survey company. Participants of the pilot test–retest were recruited through the research team members' networks, whereas those recruited by the survey company were rewarded for completing surveys, which could affect the time that they spend on each survey.

Considering an item-by-item analysis, several items did not reach the threshold of the moderate value of the reliability coefficient (between 0.41 and 0.60) (23;29). There has been considerable debate in the literature regarding the most appropriate threshold of the reliability coefficient (12). According to some authors, there is no "cut-off" point associated with an appropriate coefficient threshold (30;31). A high test–retest reliability (ICC = 0.96) was found for the item 3.2 ("What is your rough estimate [net amount] of the incurred loss of income?") in this study. This suggested that the participants' responses were highly stable for this item because the patient out-of-pocket costs could have a direct effect on their loss of income. In contrast, a lower ICC (-0.00) was obtained for the item 6.3 ("How long is the estimated waiting time experienced by your caregiver or the person accompanying you during your nonmedical consultations [e.g., massotherapy, chirotherapy, naturopath]?"). This low ICC implies either a lack of stability of the informal caregiver's waiting time or a lack of stability of the measure (32). Another possible explanation is that the respondents were not directly concerned by this item. This result is not surprising because the informal caregivers were not asked directly about the waiting time during all the medical consultations. The 3-month period chosen in this study as time frame could also lead to a memory bias because the participants did not remember the waiting time spend by their informal caregivers.

The lack of stability of some items (e.g., cost and time) over time suggested that participants responded differently between the two periods, leading to a high variability. This is more reflected with items that do not affect them directly (e.g., items specific to the informal caregivers). This situation could be due to participants having less knowledge of their informal caregiver's expenses. The reactivity effect, which is a systematic factor that can affect stability in over time, could also be a reason. Reactivity refers to a phenomenon in which respondents are sensitive to the specific issues examined by an instrument and demonstrate a change in their response at the second time (33). Generally, it occurs when respondents who are unfamiliar with the items think about questions between the two points in time and the phenomenon is particularly common when respondents are not used to answering detailed questions (33). There are reasons to believe that reactivity effect could be larger in our study for the items of cost and time. In fact, the CoPaQ is a new instrument and

participants should remember the cost or time related to the use of health services in the past 3 months. This time frame could lead to participants forgetting small ticket items and time elements. It is most likely that between the test and retest, participants remember some amounts and duration and modify their response at the retest 2 weeks later.

In the sensitivity analyses, the reliability coefficients based on the items which were specific to the patients (category 1 and 3) were higher than those that were specific to the informal caregivers (category 2 and 4), indicating a better reliability of items concerning specifically the participants. This finding suggests that if the informal caregivers could complete themselves the items that are specific to them, the reliability rates could potentially be increased.

The CoPaQ could be administered by an interviewer or be self-administered. In this study, the time spent for the self-administration (including reading the explanation and the instructions, and completion time) was about 8 min. When addressing parsimony in an instrument, it is useful to think of both length and psychometric properties (34). The length of the questionnaire and the completion time are very important because they could impact the nonresponses and the missing data rates. For example, in their comments, some participants considered that the CoPaQ was too long to complete. However, an 8-min mean completion time seems appropriate in light of the literature which reports a completion time between 5 and 10 min for a similar type of questionnaire (35). Although the time frame can easily be determined by researchers to align with their study objective, the majority (47.89 percent) of the participants favored a 3-month period recall. This period should allow them to reduce the oversight of the small ticket items.

Strengths and Limitations

The study has multiple strengths. First, we planned for and recruited more participants than the minimum size requirement to make up for any possible loss of data due to dropouts or missing data (29). The test-retest with these larger sample and diverse patient groups followed a rigorous process as described in this study. Second, the format of the CoPaQ is suitable for data analysis and the nonresponse rate in this study was relatively low during the test and the retest. The study also has some limitations. Firstly, using a survey company has disadvantages, because the participants are not like those in a clinical trial who have a vested interest in the study. Hence, there could be poor engagement. Secondly, no further psychometrics properties were evaluated apart from the content validity in previous studies (17;19) and the measure of the test-retest reliability, and such tests could be conducted in the future to increase the thoroughness of the CoPaQ. Despite this, we believe that there are several reasons why the CoPaQ can be used in the field. First, the way in which it was developed ensures content validity in that it is comprehensive and represents costs that are important to patients and their informal caregivers (17). Second, the process used to construct the CoPaQ is well established and has been used successfully in constructing specific questionnaires for patients (35).

Study Implications and Future Research

In practice, estimating the stability of test scores involves administering the same test to the same people twice in as similar conditions as possible. Once the data are collected, one correlates the scores of two test administrations. Reliability estimation yields a coefficient of stability. From this coefficient, a researcher could know how consistently people respond to the same test at different times. In

this context, the interest is in how a person's observed scores are stable over time. Kappa scores indicate how the observed values compare themselves to the expected values for each item. We note that for some items interpreted as "slight" or "fair," the level of agreement is very high, but Kappa scores may be affected by the high expected scores. Researchers may consider not only ICC and Kappa overall scores but also the level of agreements on each item to make their own judgment. The CoPaQ may be useful for researchers who would like to measure patient-related costs as well as their informal caregivers' costs. The participants were consulted about the best time frame to reduce the memory loss of the small ticket items. Their responses helped us to recommend a 3-month recall period for CoPaQ users. The validity and responsiveness tests should be evaluated in the future. A user guide is currently under development and will be made available to researchers wishing to estimate costs from the data set collected with the CoPaQ. Finally, a translation of the CoPaQ into English was foreseen.

Conclusion

Based on the findings of this study, we can conclude that the CoPaQ has a moderate reliability in terms of ICC and Kappa coefficients. The moderate test-retest reliability (i.e., approximately half of values are similar over time) observed gives additional support for the applicability of this tool in economic evaluations. The CoPaQ could be used by researchers who wish to capture out-of-pocket costs of a condition for patients and their caregivers. Additional studies including on other properties (validity, responsiveness) and a cultural adaptation could further enhance the use of the tool.

Ethics approval. The study protocol was approved by the research ethics committee of the CIUSSS de l'Estrie-CHUS (Project #2019-3102—Écosanté 2). All participants were familiarized with the purpose and principles of the study and informed about the possibility of withdrawing from the study at any stage.

Consent to participate. An online informed consent was obtained from all subjects included in the study. Data collected in the study were anonymized and do not allow identification of individual study participants.

Funding statement. This study was funded by the Quebec Support Unit of the Strategy for Patient Oriented Research (SPOR). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflicts of interest. The authors declare that they have no conflicts of interest.

Author contributions. Acquisition of data: T.G.P., L.P.C., M.L.; Analysis and interpretation of data: T.G.P., L.P.C., A.I.H., B.C., M.L.; Concept and design: T.G.P., L.P.C., M.L.; Critical revision of the paper for important intellectual content: T.G.P., L.P.C., M.L., A.I.H., B.C.; Drafting of the manuscript: L.P.C., A.I.H.; Methodology: T.G.P., L.P.C., A.I.H., M.L. All authors were involved in revising the article and approved the final manuscript.

References

1. Doherty WJ, Mendenhall TJ (2006) Citizen health care: A model for engaging patients, families, and communities as coproducers of health. *Fam Syst Health*. 24, 251.
2. Taylor J, Rutherford P (2010) The pursuit of genuine partnerships with patients and family members: The challenge and opportunity for executive leaders. *Front Health Serv Manag*. 26, 3–14.

3. **Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW** (2015) *Methods for the economic evaluation of health care programmes*. Oxford: Oxford University Press.
4. **World Health Organization** (2017) *Tuberculosis patient cost surveys: A handbook*. Available at: https://www.who.int/tb/publications/patient_cost_surveys/en/. Accessed 2021.
5. **Sánchez RÁ, Iraola AB, Unanue GE, Carlin P** (2019) TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. *Comput Meth Programs Biomed.* **181**, 104824.
6. **Canadian Institute for Health Information** (2009) The CIHI data quality framework, CIHI Ottawa (Ottawa, ON: CIHI, 2009). Available at: https://secure.cihi.ca/free_products/dq-data_quality_framework_2009_en.pdf. Accessed 2021.
7. **Richards SH, Coast J, Peters TJ** (2003) Patient-reported use of health service resources compared with information from health providers. *Health Soc Care Community.* **11**, 510–518.
8. **Lohr KN** (2002) Assessing health status and quality-of-life instruments: Attributes and review criteria. *Qual Life Res.* **11**, 193–205.
9. **De Souza JA, Yap BJ, Wroblewski K**, et al (2017) Measuring financial toxicity as a clinically relevant patient-reported outcome: The validation of the comprehensive score for financial Toxicity (COST). *Cancer.* **123**, 476–484.
10. **Rattray J, Jones MC** (2007) Essential elements of questionnaire design and development. *J Clin Nurs.* **16**, 234–243.
11. **Price LR** (2016) *Psychometric methods: Theory into practice*. New York: Guilford.
12. **Streiner DL, Norman GR, Cairney J** (2015) *Health measurement scales: A practical guide to their development and use*. Oxford: Oxford University Press.
13. **Smith S, Lamping D, Banerjee S**, et al (2005) Measurement of health-related quality of life for people with dementia: Development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technol Assess.* **9**, 1–93, iii–iv.
14. **Polit DF** (2014) Getting serious about test–retest reliability: A critique of retest research and some recommendations. *Qual Life Res.* **23**, 1713–1720.
15. **Evans C, Mertzanis P, Abetz L** (2003) Measurement strategies for indirect costs in economic evaluations. *Expert Rev Pharmacoecon Outcomes Res.* **3**, 703–716.
16. **Ness N-H, Haase R, Kern R**, et al (2020) The multiple sclerosis health resource utilization survey (MS-HRS): Development and validation study. *J Med Internet Res.* **22**, e17921.
17. **Laberge M, Coulibaly LP, Berthelot S**, et al (2021) Development and validation of an instrument to measure health-related out-of-pocket costs: The cost for patients questionnaire. *Value Health.* **24**, 1172–1181.
18. **Koo TK, Li MY** (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropractic Med.* **15**, 155–163.
19. **Poder TG, Coulibaly LP, Gaudreault M, Berthelot S, Laberge M** (2022) Validated tools to measure costs for patients: A systematic review. *Patient.* **15**, 3–19.
20. **Al-Janabi H, Flynn TN, Peters TJ, Bryan S, Coast J** (2015) Test–retest reliability of capability measurement in the UK general population. *Health Econ.* **24**, 625–630.
21. **Terwee CB, Bot SD, de Boer MR**, et al (2007) Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* **60**, 34–42.
22. **Sim J, Wright CC** (2005) The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys Ther.* **85**, 257–268.
23. **Landis JR, Koch GG** (1977) The measurement of observer agreement for categorical data. *Biometrics.* **13**, 159–174.
24. **Feinstein AR, Cicchetti DV** (1990) High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol.* **43**, 543–549.
25. **Boulan H** (2015) *Le questionnaire d'enquête: Les clés d'une étude marketing ou d'opinion réussie*: Dunod.
26. **Lerner D, Amick BC, Rogers WH**, et al (2001) The work limitations questionnaire. *Med Care.* **39**, 72–85.
27. **Duncan P, Murphy M, Man M-S**, et al (2018) Development and validation of the Multimorbidity Treatment Burden Questionnaire (MTBQ). *BMJ Open.* **8**, e019413.
28. **Sikandar MA, John V** (2015) A study to investigate test-retest reliability of two minute walk test to assess functional capacity in elderly population. *Indian J Physiother Occup Ther.* **9**, 108.
29. **Bujang MA, Baharum N** (2017) A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *Arch Orofac Sci.* **12**, 1–11.
30. **Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Costa C** (2011) The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *J Clin Epidemiol.* **64**, 264–269.
31. **Fortin F** (1994) Propriétés métriques des instruments de mesure (fidélité-validité). *Recherche en Soins Infirmiers.* **39**, 58–62.
32. **Grafton KV, Foster NE, Wright CC** (2005) Test-retest reliability of the Short-Form McGill Pain Questionnaire: Assessment of intraclass correlation coefficients and limits of agreement in patients with osteoarthritis. *Clin J Pain.* **21**, 73–82.
33. **Torkzadeh G, Doll WJ** (1991) Test-retest reliability of the end-user computing satisfaction instrument. *Decis Sci.* **22**, 26–37.
34. **Ferketich S** (1991) Focus on psychometrics aspects of item analysis. *Res Nurs Health.* **14**, 165–168.
35. **Juniper EF, Guyatt GH, Epstein RS**, et al (1992) Evaluation of impairment of health related quality of life in asthma: Development of a questionnaire for use in clinical trials. *Thorax.* **47**, 76–83.