The Canadian Journal of Program Evaluation Vol. 24 No. 2 Pages 81–102
ISSN 0834-1516 Copyright © 2010 Canadian Evaluation Society

INTERRATER RELIABILITY IN CONTENT ANALYSIS OF HEALTHCARE SERVICE QUALITY USING MONTREAL'S CONCEPTUAL FRAMEWORK

Bernard-Simon Leclerc
Agence de la santé et des services sociaux de Lanaudière
(at the time of the study)
Joliette, Québec

Clément Dassa
Université de Montréal
Montréal, Québec

Abstract: This study examines the usefulness of the Montreal Service Concept framework of service quality measurement, when it was used as a predefined set of codes in content analysis of patients' responses. As well, the study quantifies the interrater agreement of coded data. Two raters independently reviewed each of the responses from a mail survey of ambulatory patients about the quality of care and recorded whether or not a patient expressed each concern. Interrater agreement was measured in three ways: the percent crude agreement, Cohen's kappa, and the coefficient of the generalizability theory. We found all levels of interrater code-specific agreement to be over 96%. All kappa values were above 0.80, except four codes associated with rarely observed characteristics. A coefficient of generalizability equal to 0.93 was obtained. All indices consistently revealed substantial agreement. We empirically showed that the content categories of the Montreal Service Concept were exhaustive and reliable in a well-defined content-analysis procedure.

Résumé : Cette étude examine l'utilité du cadre de mesure de la qualité des services conçue à Montréal, qui a servi comme un ensemble de codes prédéfinis dans l'analyse du contenu des réponses des patients, mesurant aussi l'accord interjuge des données codées. Deux évaluateurs ont passé en revue indépendamment les réponses à un sondage postal de patients externes sur la qualité des soins et enregistré si un patient a exprimé un souci particulier. L'accord interjuge était calculé selon trois mesures : le pourcentage

Corresponding author: Bernard-Simon Leclerc, Direction du développement des individus et des communautés, Institut national de santé publique du Québec, 190, boul. Crémazie Est, Montréal, QC, H2P 1E2; <bernard-simon.leclerc@inspq.qc.ca>

d'accord, le Kappa de Cohen, et le coefficient de généralisabilité. Nous avons constaté des niveaux d'accord interjuge spécifique aux codes tous supérieurs à 96 %. Toutes les valeurs du coefficient de Kappa excédaient 0,80, à l'exception de quatre codes reliés à des traits rarement observés. Le coefficient de généralisabilité calculé se situe à 0,93. Tous les indices se sont avérés significativement en accord de façon constante. Nous avons démontré empiriquement que les catégories du contenu du concept de service de Montréal sont exhaustives et fiables dans un processus bien défini d'analyse de contenu.

INTRODUCTION

Interest in quality assessment of health-care services including patients' opinions has grown considerably in recent years. Notably, the Montreal Health and Social Service Agency has undertaken initiatives to develop quality indicators from the viewpoint of patient satisfaction with the services they received (Côté, Bélanger, Granger, & Ladeuix, 2005). The conceptual framework of service quality measurement, commonly referred to as the Montreal Service Concept, was designed primarily to be used in surveys involving Likert-scale questions for data collection and statistical analysis. Quality from a patient's point of view is defined as the extent to which their needs and expectations are met. In health services research and program evaluation, patient satisfaction is hypothesized to be an outcome measure of the quality of service as well as a predictor of subsequent health-related behaviour (Jangland, Gunningberg, & Carlsson, 2009; Montini, Noble, & Stelfox, 2008).

Qualitative data collection and analysis are now common approaches in the health services research and evaluation fields. Their prime advantage is that they result in a free, rich, and detailed description of the patient care experience. However, they are often suspected of analyst bias, subjectivity, and selective interpretation. In spite of that, the interrater reliability of any methodological process is a critical technical quality issue that needs to be considered and examined empirically, regardless of whether the study is qualitative or quantitative (Marques & McCall, 2005).

We previously produced an unpublished patients' satisfaction survey including an open-ended question to afford participants the opportunity to expand their responses to quantitative questions. Availability of these written narratives provided us with an occasion to evaluate the usefulness of the Montreal Service Concept as a predefined set of codes in content analysis of patients' responses and to quantify interrater agreement of qualitatively coded data.

CONCEPTUAL FRAMEWORK

The 2005 edition of the Montreal Service Concept identifies 44 patient expectations, grouped into three main public health service sectors of activity (relational, professional, and organizational) of overall health and social services and combined them within 12 objectives. The patient expectations, which are regularly validated in focus discussion groups, were set out as structured statements of opinion. Accordingly, quality service should provide an appropriate response to all these expectations. The Conseil québécois d'agrément (the Quebec Council on Health Services Accreditation) has formally integrated the Montreal Service Concept into its certification requirement process. The Service Concept is illustrated and described in detail in a monograph available on the website of the Montreal Health and Social Service Agency (Côte et al., 2005).

INTERRATER RELIABILITY

Interrater reliability or agreement determines the extent to which two or more independent coders or raters obtain the same result when using the same instrument. The concept addresses the consistency of the implementation of a rating system. There are a number of statistics that can be used to estimate agreement between two coders (Sicanore, Connell, Olthoff, Friedman, & Geght, 1999). In the present article, we review three current options, including the percent crude agreement for each concern, Cohen's kappa coefficient for each concern, and the coefficient of generalizability theory calculated across all concerns, which involves the portioning of total variance.

Percent crude agreement is the simplest interrater agreement measure. Considering the design of two raters each rating *n* subjects once, we can summarize the results in a two-by-two table by considering the four possible pairs of ratings. The percent crude agreement is the proportion of all subjects for whom both raters agree about the presence or absence of the code as judged by both raters. Miles and Huberman (1994) suggested that interrater agreement in qualitative data analysis should approach or exceed 90%. However, proportion of crude agreement has been criticized since it fails to account for any agreement due to chance alone. Consequently, it can lead to inflated true levels of reliability.

Cohen proposed a method, the Cohen's kappa statistic (κ), to counteract this problem of agreement caused by chance. The computation subtracts the proportion of agreement that could be expected by chance alone from the observed proportion of agreement (Fleiss, 1981). The maximum possible value of kappa is 1, indicating perfect agreement. Inversely, values of kappa between 0 and -1 signify that agreement is less than can be attributed to chance. According to the classification interpretative guidelines suggested by Landis and Koch (1977), kappa values of 0.40 to 0.60 represent moderate agreement, 0.60 to 0.80 substantial agreement, and 0.80 to 1.00 almost perfect agreement. However, although kappa is widely accepted as an appropriate statistic, there are two situations that adversely influence the magnitude of the values and may confound their interpretation: the prevalence of the trait (e.g., infrequently found codes) and the pattern of disagreement between raters (e.g., unbalanced marginal totals in contingency tables) (Feinstein & Cicchetti, 1990; Graham & Jackson, 1993; Guggenmoos-Holzmann & Vonk, 1998; Gwet, 2002; Hoehler, 2000; Lantz & Nebenzahl, 1996).

Coefficients based on the theory of generalizability is an alternative approach for measuring agreement that does not present as many weaknesses as the crude and chance-corrected agreement indices. It represents the most comprehensive and complex way of assessing reliability. The ideas of the generalizability theory have been addressed extensively in the statistical literature and have become increasingly popular in various fields of science (Donnon & Paolucci, 2008; Kan, 2007; Lakes & Hoyt, 2009; O'Brian, O'Brian, Packman, & Onslow, 2003; Odegård, Hagtvet, & Bjørkly, 2008; Vangeneugden, Laenen, Geys, Renard, & Molenberghs, 2005; Wasserman, Levy, & Loken, 2009). The coefficients of generalizability (ρ^2), also called intraclass correlation coefficients, are a generic class of reliability coefficients. They might be derived from a specific formula depending on the experimental plan and calculated from an analysis of variance approach (Crocker & Algina, 2008; O'Brien, 1990). The values of generalizability coefficients range from 0, which indicates no agreement, to 1, indicating perfect agreement among raters. Values are regarded as acceptable if they equal or surpass 0.8 (Landis & Koch, 1977; Nunnally, 1978).

METHODS

Data Collection and Study Subjects

A French-language anonymous mail survey among outpatients requiring diabetes, intravenous antibiotherapy, or post-surgical care was conducted within 3–4 weeks after discharge from hospital. The survey involved a quantitative measurement approach, which also included a complementary open-ended question. Closed-question topics included satisfaction about preparations made in hospital for discharge; links made by hospital staff members with post-discharge caregivers; preparation for self-care at home; perceptions of the quality of the ambulatory basis care provided by caregivers from the public regional hospitals, community services, and medical clinics; and satisfaction about contact with staff members they worked with. To ascertain additional self-reported comments on patients' personal experiences with health professional and medical care services, they were asked: "Do you have any critical comments or suggestions to address to us in order to improve the quality of care and services offered in your region?"

Coding Procedures

We used a common content analysis procedure (Carey, Morgan, & Oxtoby, 1996). Each patient's response was entered exactly as written directly into a computer file and marked with identification numbers. No criteria or codes were determined before the evaluation project. The principal author began by reading all the available respondents' comments and then compiling an initial working draft of the codebook as each new idea was encountered. This exercise revealed that a large part of the content of what patients said could be analyzed and summarized according to the Montreal Service Concept. We thus decided to adapt the Service Concept by adding some codes and supplementary definitions to existing codes, and to use them as a final set of reference codes. The final codebook contained 17 hypothesized all-inclusive and mutually exclusive codes, including one code labelled *acknowledgment* for all general acknowledgements and comments with comparison to earlier experiences and another labelled *unclassified* for any comments that have nothing to do with the service quality and patient satisfaction. Appendix A (1–4) shows the final list of codes and their definitions.

Two graduate-level research assistants were integrated into the project as coders, after the final codebook was entirely developed. They were trained by the first author. Training included reading available promotional material prepared by the Montreal Health and Social Service Agency about the Service Concept and reports of performed surveys that used it. To train the coders in the use of the codebook, we selected a random subset of responses given by nearly 10% of the participants (n = 68) in the sample and they independently attached codes to segments of text, such as a word, phrase, sentence, or paragraph. We compared the set of codes that each rater assigned to the response of each participant and discussed the reasons for their disagreements at each step.

The next step was to ensure that a different rater could independently replicate the other's work using the same instructions. The same two raters independently coded the comments from the remaining 90% of the respondents (n = 573) in the sample following the chronological order of identification numbers. Because many responses contained multiple dimensions, the number of codes assigned to comments of each individual varied, but a specific code could be used only once per patient response.

Statistical Procedures

Interrater reliability was first assessed through the use of indices of agreement. The overall interrater agreement was evaluated from two-by-two cross-classification tables showing the presence or absence of each separate code as judged by both raters across the 573 responses. Proportions of agreement (the number of times that both raters assigned the code added to the number of times both raters did not assign the code to a patient's comment) were compared to the proportion of the two discordant categories using the McNemar chi-square test for related variables (Fleiss, 1981). The equivalency between marginal homogeneity and symmetry tests within a two-by-two contingency table is demonstrated in Appendix B. The differences in the number of times a code was assigned by one but not by the other were also analyzed in order to examine whether a rater systematically assigned a specific code more often than the other rater. Statistical differences were estimated through the Wilcoxon matched-pairs signed rank test because the data were skewed. The significance level for all two-sided tests was fixed to 0.005 given that multiple comparisons were examined. Cohen's kappa statistic, which adjusts for agreement caused by chance, was also calculated (Fleiss, 1981).

Interrater reliability was next assessed by partitioning the total variance into relevant components of interest. The balanced design under consideration is the following: the same raters independently rated each subject's script on each of a set of codes. Labelling subjects as p (with n'_p = 573), raters as facet J (with n'_j = 2 raters), and codes as facet I (with n'_i = 17 codes), it was stated that rater facet is random and code facet is fixed. This is a fully crossed two-facet in the terminology of generalizability theory, denoted as (p × i × j). As shown in Figure 1, seven components of variance in terms of expected mean square (EMS) can be separately estimated from this design (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). They have been made by using some algebraical manipulation of the mean square (MS) quantities for main effects of p, i, and j, plus MS for interactions pi, pj, and ij, plus MS for the residual (pij) generated from the output of three-way analysis (see Table 1 for formula).

Table 1
Selected Equations for Expected Mean Square (EMS) in a Crossed Two-Facet Design (*p* × *i* × *j*)

Source of variation	EMS
p	$\sigma^2_e + n_i n_j \sigma^2_p + n_j \sigma^2_{pi} + n_i \sigma^2_{pj}$
i	$\sigma^2_e + n_j n_p \sigma^2_i + n_j \sigma^2_{pi} + n_p \sigma^2_{ij}$
j	$\sigma^2_e + n_i n_p \sigma^2_j + n_i \sigma^2_{pj} + n_p \sigma^2_{ij}$
pi	$\sigma^2_e + n_j \sigma^2_{pi}$
pj	$\sigma^2_e + n_i \sigma^2_{pj}$
ij	$\sigma^2_e + n_p \sigma^2_{ij}$
Residual (pij)	σ^2_e

Note. From Crocker & Algina, 2008

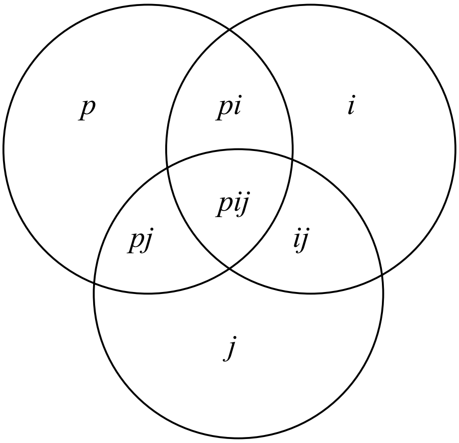
The residual variation combines the person-rater-code interaction with variation from unidentified sources. The appropriate generalizability coefficient (ρ^2) for the data at hand is then estimated by using (Crocker & Algina, 2008; O'Brien, 1990):

$$\rho^2 = (\sigma^2_p + \sigma^2_{pi} / n'_i) / (\sigma^2_p + \sigma^2_{pi} / n'_i + \sigma^2_{pj} / n'_j + \sigma^2_e / n'_i n'_j).$$

In our fieldwork procedure, as assignment of codes to a same patient's response differs between raters, interrater disagreement increases, error constitutes an increasing portion of what is observed, reliability therefore decreases, and ρ^2 tends to its minimum value of zero. On the other hand, as interrater agreement increases, error constitutes a decreasing portion of what is observed, reliability therefore increases, and ρ^2 approaches its maximum value of unity.

Coding, organization of the database, and statistical analyses were made by using Microsoft Access, Genova (Brennan, 2001), and the Statistical Package for Social Sciences (Mushquash & O'Connor, 2006).

Figure 1
Schematic Representation of Components of Variance for a Crossed Two-Facet Design



RESULTS

The respondents were mainly women (58%); 28% were aged 18–34 years, 34% aged 35–54 years and 19% over 65 years; and 76% had completed 12 years of education or less. Sixty percent were concerned with ambulatory care related to surgery, 21% to intravenous treatment, and 11% to diabetes; the others did not provide this information. Out of 1188 returned questionnaires, 641 answered the open-ended question, from which 68 were used for training and 573 for testing. Comments varied widely in length, content, and complexity between respondents. Obviously, given that patients were invited to produce any comments or suggestions about the ambulatory care and services that they received in order to improve their quality, it is not surprising to note that comments voiced by the majority were essentially concerned with some dissatisfaction or negative characterization of nearly all of the examined dimensions of service (results not presented).

Final assignment from respondents yielded 1001 codes for coder A and 931 codes for coder B. A mean of 1.8 codes was used per individual response by coder A, with a range of 1 to 8 codes per response (mode: 1; median: 1), and a mean of 1.6 codes by coder B, with a range of 1 to 6 codes per response (mode: 1; median: 1). Overall, coder A thus tended to significantly assign a higher number of codes to a respondent than did coder B ($p < 0.001$).

Content analysis of patients' open-coded responses and interrater agreement results are displayed in Table 2. Following is an example of a response to illustrate the assignment of codes by the two raters. Each of them has attached the same *courtesy* and *simplicity* codes to this segment of response.

> If the care-giver had a smile it would be nicer. Going to the hospital is already stressful enough, so if the people could smile more it would be easier. The nurses are always using big words. If we haven't studied the same things how are we supposed to know what they're talking about?

The number of codes assigned looks relatively similar. However, coder A tended to assign the *empathy* code more often than coder B (13.1% against 10.1%, $p < 0.001$). This marginal homogeneity test is equivalent to saying that the proportions of assignments of *empathy* code by A but not by B (A^+B^-) was significantly higher than assignments of the same code by B but not by A (A^-B^+).

In general, the higher percentage of times that the two raters agreed (sum of A^+B^+ and A^-B^-) reveal that the two raters used the majority of the codes in the same way across the 573 responses. Total disagreement (sum of A^+B^- and A^-B^+) appeared slightly highest for *empathy* (3.6%), *accessibility* (3.2%), and *acknowledgement* (3.8%) codes. Kappa values of 0.81 or greater were achieved for almost all codes. Only *privacy* (0.67), *promoting self-sufficiency* (0.69), *solidarity* (0.50), and *simplicity* (0.71) codes (for which, however, the prevalence of the underlying trait is very small) had lower kappa values.

Table 3 presents the variance components from the generalizability theory analysis of the data. Further examination reveals that the variance due to the subject by the rater interaction is only 1% of the sum of the relevant observed variance components. The appropriate generalizability coefficient, ρ^2, equals 0.93.

DISCUSSION

The methodological issue addressed by this article is the reliability of coding responses, generated by an open-ended question, by two different raters. To our knowledge, this is the first study to use the Montreal Service Concept as a predefined set of codes in content analysis of healthcare service quality. It seems that the Montreal Service Concept, after slight adaptation, may be useful and applica-

Table 2
Interrater Agreement Results of Responses to an Open-Ended Question from a Survey of Service Quality and User Satisfaction

Dimension (code)	Percentage of assignment		Percentage of agreement[a]		Percentage of disagreement[a]		Kappa
n = 573	Coder A	Coder B	Total[b]	A^+B^+	A^+B^-	A^-B^+	
Respect	6.5	5.4	97.9	4.9	1.6	0.5	0.81
Privacy	0.3	0.2	99.8	0.2	0.2	0.0	0.67
Empathy	13.1	10.1***	96.3	9.8	3.3	0.3***	0.82
Courtesy	11.2	10.3	98.4	9.9	1.2	0.3	0.92
Realiability	13.1	11.5	97.0	10.8	2.3	0.7	0.86
Self-sufficiency	3.7	3.3	97.9	2.4	1.2	0.9	0.69
Reassurance	10.5	8.7	97.2	8.2	2.3	0.5	0.84
Solidarity	0.3	0.3	99.7	0.2	0.2	0.2	0.50
Simplicity	2.8	2.1	98.6	1.7	1.0	0.3	0.71
Continuity	16.9	17.3	97.2	15.7	1.2	1.6	0.90
Accessibility	11.9	10.5	96.9	9.6	2.3	0.9	0.84
Rapidity	13.8	13.6	98.8	13.1	0.7	0.5	0.95
Comfort	7.2	7.0	99.1	6.6	0.5	0.3	0.93
Support	4.2	3.5	98.6	3.1	1.0	0.3	0.81
Sufficiency	13.3	13.1	97.0	11.7	1.6	1.4	0.87
Acknowledgment	40.3	38.9	96.2	37.7	2.6	1.2	0.92
Unclassified	5.8	6.6	98.8	5.6	0.2	1.0	0.90

[a] A^+, B^+, A^-, B^-: assignments of the code by coders A and B, simultaneously. [b] Sum of positive (A^+B^+) and negative (A^-B^-) intercoder agreements.
***$p \leq 0.005$.

ble to other investigators for classifying and analyzing the content of responses about service quality and patient satisfaction surveys (Côté et al., 2005). The set of qualitative codes drawn from this framework appeared sufficiently exhaustive to include nearly all expectations and comments voiced by our patients' sample. Only 5.8% to 6.6% of the patients' comments couldn't be classified into one or other of the dimensions.

Using three ways to approach the question of interrater reliability, we showed that agreement between raters for their assignment of predetermined codes to key concepts was consistent. Obviously, using any cutoff point to judge what is "sufficiently high" is an arbitrary decision. Nevertheless, the triangulation of statistical methods enhances our confidence in the conclusion that the set of codes used was used similarly by the two raters. We note just one drawback in our results. Agreement computed from kappa statistics appeared poorer for *privacy*, *promoting self-sufficiency*, *solidarity*, and *simplicity* codes, for which, however, the prevalence of the underlying trait is very small. In such a situation we really do not have sufficient information to judge agreement very well, and a kappa coefficient may not be appropriate (Hoehler, 2000).

Table 3
Summary ANOVA Results for the Data

Source of Variation	SS	DF	MS	EMS	%
Persons (*p*)	62.11	572	0.109	0.00296	3
Codes (*i*)	151.09	16	9.443	0.00762	9
Raters (*j*)	0.25	1	0.252	0.00003	0
pi	1428.21	9152	0.156	0.06861	77
pj	4.51	572	0.008	0.00046	1
ij	0.38	16	0.024	0.00002	0
Residual (*pij*)	93.85	9152	0.010	0.00965	11

Note. The term *analysis of variance* comes from the process of partitioning the total variability into parts. The data were collected by using a crossed two-facet design (*p* × *i* × *j*). SS: sum of squares; DF: degrees of freedom; MS: mean square; EMS: expected mean square.

Reliability is a fundamental concern for ensuring rigour of qualitative studies in a positivist research paradigm. It is surely impossible for any researcher to be completely objective, but it is each researcher's responsibility to minimize the opportunity for problems. Unfortunately, researchers rarely attempt to check and quantify problems of reliability (Wiggins, 2004). In spite of that, the concept of reliability emerges implicitly in descriptions of procedures for carrying out the analysis of qualitative data. First, qualitative methodologists often stress the transparency of their technique by documenting all steps of the process in detail. This implies an assumption that collected data can be checked by independent researchers to reject or sustain the original interpretations. Second, qualitative methodologists frequently report carrying out the analysis together as a team activity. This latter practice is then believed to produce better analysis than one made by a lone researcher, given that results will be enhanced if one view is tempered by another or negotiated between two.

Barbour (2001) added in this sense that "what is ultimately of value is the content of disagreements and the insights that discussion can provide for refining coding frames" (p. 1116). The application of the method in our study revealed its potential usefulness in helping to identify discrepancies between raters and areas where consensus meetings or additional training would be desirable. Interrater agreement could help us to pinpoint precisely which codes gave the greatest problems. Such problematic codes can be identified in two ways: those with the greatest percentages of total disagreement and those with greater differences in the manner they were used by the two raters. Here, it might make sense to take such problems into consideration, especially when both raters are sometimes indecisive on how to classify responses and when, as in many of the cases, raters allowed more than one code per individual response. This guesswork, however, may differ from one dimension of service to another because those based on clearness criteria may be easier and other dimensions more difficult to classify.

Our article addressed the question of the use of quantitative parameters in qualitative research to evaluate studies, and more especially the use of counts of codes and percentages based on codes. The intent of this article is mainly expository. Its purpose was to describe reliability, not to report the results of the service quality and patient satisfaction. However, readers must keep in mind that the interpretative basis in qualitative analysis is more than just a history of those codes. The analysis must go further than this point of examining tables of counts in order to learn from the qualitative data and draw useful conclusions.

There are some limitations in our study that limit excessive optimism. First, the development of our set of codes was made or adapted by one of our researchers based on all the material. Thus, it may represent overfitting of data. It might have been a strength to develop the coding scheme on a subset of the data, and apply it on the rest of the data. In order to judge the real appropriateness of the Service Concept in a similar context, it should be validated by using a new sample. Second, the test presented a relatively simple scenario, among a multitude of other possibilities, of qualitative research in the "real world": a single researcher established the coding scheme (without validation by another senior researcher endeavouring to do the same thing) in the context of a close-ended instrument that framed respondents' ideas in a predefined way; then two (and only two) other coders applied the coding scheme. We do not know what would have been the results in a more complex and stringent context.

Anyone carrying out reliability analyses should be aware of the discussed issues. Any interpretation regarding good/bad reliability should be cautiously adapted unless, as in our study, all methods point to the same conclusion. In conclusion, we empirically showed that the content categories were mutually exclusive, exhaustive, and reliable in a well-defined content-analysis procedure.

ACKNOWLEDGEMENTS

We thank Bruce Charles Bezeau, who contributed to the linguistic revision of the manuscript. Special thanks to Jan Brogger from the University of Bergen in Norway, Constantine Daskalakis from Thomas Jefferson University in the USA, and Normand Peladeau from Provalis Research in the Canadian province of Quebec, for their judicious comments on an earlier draft.

REFERENCES

Barbour, R.S. (2001). Checklists for improving rigour in qualitative research: A case of the tail wagging the dog? *British Medical Journal, 322*(7294), 1115–1117.

Brennan, R.L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Carey, J.W., Morgan, M., & Oxtoby, M. (1996). Intercoder agreement in analysis of responses to open-ended interview questions: Examples from tuberculosis research. *Cultural Anthropology Methods, 8*(3), 1–5.

Côté, L., Bélanger, M., Granger, R., & Ladeuix, C. (2005). *Assessing to improve: The essence of quality services. The service concept: A tool to assess user satisfaction based on user expectations.* Montreal: Agence de la santé et des services sociaux de Montréal. <http://www.cmis.mtl.rtss.qc.ca/pdf/publications/isbn2-89510-296-1.pdf>

Crocker, L.M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Donnon, T., & Paolucci, E.O. (2008). A generalizability study of the medical judgment vignettes interview to assess students' noncognitive attributes for medical school. *BioMed Central Medical Education 8*, 58.

Feinstein, A.R., & Cicchetti, D. (1990). V. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York, NY: Wiley.

Graham, P., & Jackson, R. (1993). The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology, 46*(9), 1055–1062.

Guggenmoos-Holzmann, I., & Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine, 17*(8), 797–812.

Gwet, K. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods For Inter-Rater Reliability Assessment, 2*, 1–9.

Hoehler, K.K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology, 53*, 499–503.

Jangland, E., Gunningberg, L., & Carlsson, M. (2009). Patients' and relatives' complaints about encounters and communication in health care: Evidence for quality improvement. *Patient Education and Counselling, 75*(2), 199–204. doi:10.1016/j.pec.2008.10.007

Kan, A. (2007). Effects of using a scoring guide on essay scores: Generalizability theory. *Perceptual and Motor Skills, 105*(3 Pt 1), 891–905.

Lakes, K.D., & Hoyt, W.T. (2009). Applications of generalizability theory to clinical child and adolescent psychology research. *Journal of Clincal Child and Adolescent Psychology, 38*(1), 144–65.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lantz, C.A., & Nebenzahl, E. (1996). Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology, 49*(4); 431–4.

Marques, J.F., & McCall, C. (2005). The application of interrater reliability as a solidification instrument in a phenomenological Study. *The Qualitative Report, 10*(3), 439–462.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Montini, T., Noble, A.A., & Stelfox, H.T. (2008). Content analysis of patient complaints. *International Journal for Quality in Health Care, 20*(6), 412–420.

Mushquash, C., & O'Connor, B.P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542–547.

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

O'Brian, N., O'Brian, S., Packman, A., & Onslow, M. (2003). Generalizability theory I: Assessing reliability of observational data in the communication sciences. *Journal of Speech, Language, and Hearing Research, 46*(3), 711–7.

O'Brien, R.M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods & Research, 18*(4), 473–504.

Odegård, A., Hagtvet, K.A., & Bjørkly, S. (2008). Applying aspects of generalizability theory in preliminary validation of the Multifacet Interprofessional Collaboration Model (PINCOM). *International Journal of Integrated Care, 8*, e74.

Sicanore, J.M., Connell, K.J., Olthoff, A.J., Friedman, M.H., & Geght, M.R. (1999). A method for measuring interrater agreement on checklists. *Evaluation and the Health Professions, 22*(2), 221–234.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics 61*(1), 295–304.

Wasserman, R.H., Levy, K.N., & Loken, E. (2009). Generalizability theory in psychotherapy research: The impact of multiple sources of variance on the dependability of psychotherapy process ratings. *Psychotherapy Research, 19*(4–5), 397–408.

Wiggins, G. (2004). The analysis of focus groups in published research articles. *Canadian Journal of Program Evaluation, 19*(2), 143–64.

Appendix A1
Dimension in the Relational Sector from the Adapted Montreal Service Concept

This sector, which concerns relationships with users, covers humane treatment and aspects of service that users consider extremely important. These dimensions depend very much on the behavior of each staff member in the institutions in the network.

Dimension	*Definition*	*Expectation (Statements)*
Respect	Respect means the respect and consideration to which everyone is entitled.	• You are treated with respect • Your choices are respected and you are not forced to participate in particular activities • Staff not only deal with your immediate problem, but treat you as a whole person • Staff consider your culture and lifestyle • Everyone is treated equally, without any form of discrimination
Privacy	Privacy means that individuals receive personalized, confidential and safe treatment. This requires an environment that is physically comfortable, and attentive, discreet behaviour.	• You are not asked questions that have nothing to do with your problem • Staff deal with you in a discreet way • Your case is handled confidentially • Your physical privacy is respected by the institution
Empathy	Empathy involves understanding others' feelings. This implies active listening	• Staff takes the time to deal with you • Staff is available to meet your expectations and needs • Caregivers listen to you attentively and take you seriously • Staff shows consideration for members of your family who are helping you • Staff understands you properly
Courtesy	Courtesy involves polite, kind, and considerate behaviour.	• Staff is welcoming, amiable, and cheerful

Appendix A2
Dimension in the Professional Sector in the Codebook Adapted from the Montreal Service Concept

This sector, which covers the provision of professional services, refers to specific professions, expertise, and ways of meeting basic needs through treatment.

Dimension	*Definition*	*Expectation (Statements)*
Reliability	Reliability means the assurance that everything will go smoothly within known, reasonable limits (e.g., time and environment), as promised either implicitly or explicitly. This involves competence, consistency, thoroughness, and transparency.	• Appointments are kept on time • You are able to assert your rights if you are not satisfied with the services offered • Staff is competent • You are given the results of any examinations and tests you have
Promoting self-sufficiency	Promoting self-sufficiency is everything that helps to increase an individual's autonomy and capacity to take initiatives, assume responsibilities, and exercise leadership in his or her own regard.	• Caregivers present the various choices available • Caregivers help you find your own solution • No one makes decisions on your behalf • Caregivers advise you on how to avoid a recurrence of the problem • Caregivers advise you on what to do and what not to do once at home
Reassurance	Reassurance is the ability to calm and reassure a person and make him or her feel secure.	• Caregivers take the time to give a full step-by-step explanation of what is happening • Caregivers answer all your questions
Solidarity	Solidarity involves everything that encourages the patient to reach out to his or her family and community to get them involved in solving problems, to a greater or lesser degree.	• The institution encourages you to get support from family, friends, and community groups • You are put in contact with other people or associations of people who have experienced similar problems

Appendix A3
Dimension in the Organizational Sector in the Codebook Adapted from the Montreal Service Concept

This sector, which involves service organizations, covers the environment, contacts, and relationships between users and caregivers. The context may be helpful or not so helpful, with a certain comfort level: service may be more or less accessible, and systems, policies, and procedures may or may not be designed to ensure rapid service, continuity, and facility of service.

Dimension	*Definition*	*Expectation (Statements)*
Simplicity	Simplicity means services that are easy to use, easy to understand, and flexible in various circumstances. "Easy" applies to both the people involved (whose behaviour should be natural, spontaneous, and unpretentious) and the things involved (which should be easy to understand and use).	• There are not too many formalities in your dealings with the institution • You are spoken to in a way you can understand • You feel that you can choose your caregivers or switch if things do not go well • You are given all the information you need on where to go, what to do, and what not to do
Continuity	Continuity is the assurance that users will receive full treatment, without any interruptions in case management, responsibilities, or information.	• You feel that there is continuity and a good flow of information between various health resources, caregivers, and you • The same caregivers see you on each visit • You feel that you were kept at the hospital long enough to recover • You feel that you may see the doctor again if necessary • You feel that you can receive a medical or nursing follow-up at home if necessary
Accessibility	Accessibility means that the institution is accessible in terms of geography, physical facilities, schedules, and culture.	• The institution is easily accessible via public transit or has parking nearby • The institution provides information on all the services it offers

(continued)

Accessibility *(continued)*		• The information you are given is adapted to your culture and language • The institution is open at convenient times (weekdays, weekends, at lunch hour, evenings, and/or night) • The institution is geographically convenient • The necessary costs for care and services are affordable (transportation, parking, drugs, physiotherapist, etc.) • It is possible to see a doctor other than in the emergency department if you want
Expediency	Expediency means how long it takes to get a response to a request, depending on the client's expectations and needs.	• You can make an appointment to see a professional quickly • If you don't have an appointment, you don't have long to wait • The results of examinations and tests are obtained quickly
Comfort	Comfort is the feeling of well-being users get from being in sanitary, clean, and orderly surroundings, with equipment that can be adapted to various situations.	• The atmosphere in the institution is pleasant • The premises and equipment are clean • You are satisfied with the tranquility of the hospital room • The food served is good and appetizing • You are happy with the schedule of meals
Support	Support refers to any physical and psychological help that a client could need to maintain their autonomy	• You can have help to get around the institution if you want • You can have support and help with your activities of daily living and domestic chores • You feel that one doesn't assume that you can count on the support of family and friends

Sufficiency	Sufficiency means the capital investment in the health and service network is as much as is needed. The adequate amount and diversity of services are put at the disposal of patients to respond to their needs and expectations.	• You can have the amount and diversity of services and resources to respond to your needs and expectations, within reason

Appendix A4
Unspecific Dimension in the Codebook Adapted from the Montreal Service Concept

This sector covers all aspects which cannot be classified in the relational, professional or organizational sectors of quality service.

Dimension	*Definition*	*Expectation (Statements)*
Acknowledgement	Acknowledgement includes all general comments and acknowledgments expressed about an institution or a caregiver without any specific details, and all general comments with comparison to experiences in other institutions, regions or time period.	• You thank or congratulate an institution or somebody on doing something. • You feel that the current quality of service is better or worse than what you experienced earlier in the same place, or in other institutions or regions.
Unclassified	This code is assigned when no other classification is appropriate for any segment of the response.	• Any comments that have nothing to do with the service quality and user satisfaction.

Appendix B
Equivalency Between Marginal Homogeneity and Symmetry Tests Within a Two-by-Two Contingency Table

If one cross-classifies the data from two raters as:

	Rater B		
Rater A	a	b	a+b
	c	d	c+d
	a+c	b+d	N

Then, testing $(c+d)/N = (b+d)/N$ [1] is equivalent to testing $b/N = c/N$ [2]. This is easy to show because $(c+d)/N = (b+d)/N$ is equivalent to $c/N + d/N = b/N + d/N$ and, because the d/N cancels out from both sides, this is equivalent to $c/N = b/N$. Sometimes, [1] is called marginal homogeneity and [2] is called symmetry.

Bernard-Simon Leclerc, Ph.D., is an epidemiologist and health interventions analyst and evaluator at the Institut national de santé publique du Québec. He is also clinical assistant professor in the Département de médecine sociale et préventive of the Université de Montréal and associate researcher at the Centre de recherche de l'Institut universitaire de gériatrie de Montréal. At the time of the development of this study, he was employed at the Agence de la santé et des services sociaux de Lanaudière.

Clément Dassa, Ph.D., is a full professor, senior statistician, and psychometric specialist in the Département de médecine sociale et préventive of the Université de Montréal as well as research member of the Groupe de recherche interdisciplinaire en santé and the new Institut de recherche en santé publique de l'Université de Montréal. He acts as a consultant to groups that require complex multidimensional statistical analysis.