Psychometrics Properties of the Questionnaire Surveys: Validity, Reliability and Feasibility

Hana Zach[1]
Department of Statistics and Operation Research, Slovak University of Agriculture in Nitra

Psychometrics Properties of the Questionnaire Surveys: Validity, Reliability and Feasibility. All questionnaires, regardless of what they measure, must demonstrate good performance with regard to psychometric properties. Psychometrics is a branch of survey research that has developed methods how to quantify errors in measurement because no matter how well is the questionnaire prepared, observed data bears except of desired true data also measurement errors. You should distinguish notion psychometrics from psychometrics used in sociological practice. There is no Slovak equivalent for psychometrics (as a branch of survey research) at the moment, although term psychometrics is common abroad and comprehended correctly. But we can also say we are introducing methodological requirements for data collected through the questionnaire surveys that should be satisfied prior to any statistical manipulation. Řehák (1998b) noticed the quality of data determinate the quality of results. In general, questionnaires collecting research data should satisfy properties of validity, reliability and feasibility. Criteria and standards in psychometrics are not defined as strict rules, but rather as ranges within you would expect certain indicators to fall if the questionnaire is operating correctly. Also the strict limit showing which and how many tests must be used to proof the quality of questionnaire is missing. According to the most of reviewed literature authors use the same set of tests that vary only in number of tests according to which psychometrics areas were tested. Methodology, how to test your quality of data collected through the questionnaire survey, is demonstrated on empirical Slovak data achieved through the Multi-country Survey Study (MCSS) questionnaire.
Sociológia 2005 Vol. 37 (No. 3: 275-291)

One ways of collecting information is using questionnaire surveys. No matter how well is the questionnaire prepared, observed data bears two basic attributes – true data and measurement errors composed of systematic measurement error and random measurement error. There are known many sources of both measurement errors, such as methodological factors, sampling strategies or irregularities in data collection. They may also be influenced by respondent characteristics as age, education, income, expectations or self-interest. In order to produce reliable survey outcome, the attempt of all researchers is to minimise measurement errors. If the questionnaire is free of random measurement error we say that it is highly reliable. If the questionnaire is free of systematic measurement error we say that it is valid (De Vellis, R., 1991). Thus, before any statistical manipulation, evaluation and final interpretation of achieved results gained through the surveys, questionnaire should be tested to address the necessary quality. Because of that requirement, psychometrics, a branch of survey research, has developed methods how to quantify errors in measurement. The aim of presented article is to step-by-step provide you with the basic psychometrics methodology, relevant also to sociological surveys.

Material and Methods

Any questionnaire, no matter if it is focused on sociological, economics or health topics must satisfy psychometric properties. The basic scheme of psychometric testing is exactly the same. Nevertheless there are a lot of authors testing psychometrics in sociological area (Choi, N., 2004; Snelgrove,S. – Slater, J., 2003; Johnson, L. M., et al. 2002; Terry, R., 2000; Speer, P. W. – Peterson, N. A., 2000; Gentry, M. – Owen, S. V., 2004 and many others) we will present the psychometrics methodology on health data. However, because of the journal focus, we will indicate in brackets relevant to particular statistical method whose "sociological" author used that test to verify the data quality. For example Surynek et al. (2002) mentioned validity and reliability testing in sociological surveys is necessary, unfortunately, they do not suggested statistical methods how to produce the results. Řehák and Řeháková (1986) focused on statistical methods used in sociological surveys evaluation, they also pointed validity and reliability testing is inevitable, but they did not concentrated on psychometric properties. Validity and reliability testing in sociological surveys, what is actually the part of psychometrics, was presented in Řehák (1998a, 1998b) and Řehák et al. (1998)

[1] Ing. Hana Zach, Department of Statistics and Operation Research, Faculty of Economics and Management, Slovak University of Agriculture in Nitra, Tr. A. Hlinku 2, 949 67 Nitra, Slovak Republic, e-mail: Hana.Zach@fem.uniag.sk

suggesting to use the same methods we are presenting in text below (such as test – retest, Cronbach`s alpha, factor analysis, inter – item correlations).

To present the psychometrics methodology we tested the Slovak dataset collected through the Multi-country Survey Study on Health and Health System's Responsiveness (MCSS) collected by World Health Organisation (WHO) in 2001 in Slovak Republic available to download from the WHO official web page. Because of the great survey extent we focused only on Responsiveness module – ambulatory care. Appendix 1 presents the question wording only for that part of evaluated questionnaire. More information about the survey background, questionnaire structure, wording for the whole questionnaire or data can be obtained from the WHO official web page (www.who.int).

In brief, we used the most common psychometric properties described in recent literature (please see all Bibliography below). Questionnaire validity was tested by factor analysis the principal factor method and Cronbach`s alpha. Reliability was reviewed by test – retest Kappa and the feasibility according to response and missing rates. Explained in more details below. We used Stata software for all statistical computations.

Results and Discussion

Validity is concerned with the extent to which the questionnaire actually measures what it is supposed to measure. Validity testing is an attempt to support the measurement assumption that specific items in the questionnaire are clear representations of the concept under study. Basic validity concept comprises of different validity types, such as face validity, content validity, criterion validity (concurrent or predictive) or construct validity (Litwin, M. S., 1995; Bland, J. M. – Altman, D. G., 2002).

Construct validity is of greatest interest of many researchers (Jenkinson, C. et al. 2003; Haddad, S., et al., 1998; Haddad, S. et al., 2000; Westaway, M. S. et al., 2003; Gigantesco, A. et al., 2003; Steine, S. et al., 2001; Verho, H. – Arnetz, J. E., 2003; McGuiness, C. – Sibthorpe, C., 2003; Labarere, J. et al., 2001; Li, L. et al., 2003; Paddock, L. E. et al., 2000; Salomon, L. et al., 1999; Gasquet, I. et al., 2004; Choi, N., 2004) and it is a theoretical measure that shows how meaningful the survey questionnaire is when in practical use. Measurement of construct validity involves a statistical evaluation of survey data by computing how highly the items are correlated. If such correlations were low, it would suggest that all items are not measuring the same construct. In other words the data will show a multidimensional structure as opposed to a desired unidimensional structure. Construct validity is usually assessed by factor loading, such as the use of factor analysis (Jenkinson, C. et al., 2003; Haddad, S. et al., 1998; Haddad, S. et al., 2000; Westaway, M. S. et al., 2003; Gigantesco, A. et al., 2003; Steine, S. et al., 2001; Řehák, J., 1998a; Řehák, J. et al., 1998), confirmatory factor analysis (Verho, H. – Arnetz, J. E., 2003; Gentry, M. – Owen, S. V., 2004), exploratory factor analysis (Verho, H. – Arnetz, J. E., 2003; Westaway, M. S. et al., 2003; Snelgrove, S. – Slater, J., 2003), ANOVA (Westaway, M. S. et al., 2003; Verho, H. – Arnetz, J. E., 2003; Salomon, L. et al., 1999) or principal components analysis (McGuiness, C. – Sibthorpe, C., 2003; Labarere, J. et al., 2001; Li, L. et al., 2003; Westaway, M. S. et al., 2003; Paddock, L. E. et al., 2000; Salomon, L. et al., 1999; Blazeby, J. M. et al., 2004; Speer, P. W. – Peterson, N. A., 2000).

We can use the generic term factor analysis referring to statistical methods such as exploratory factor analysis (EFA), principal components analysis (PCA) or confirmatory factor analysis (CFA) that are used to reduce a large amount of statistical data to a manageable and meaningful size. Confirmatory Factor Analysis (CFA) is used to confirm or reject a hypothesis about the underlying dimensionality of a construct. Actually, we want to confirm our theoretical assumption that a set of questions measuring people's experience with the health systems can be expressed in a small number of responsiveness domains. We assume that there are seven domains of responsiveness in relation to experience of ambulatory care services. In our case the purpose of the factor analysis is to test whether each set of questions can explain these seven responsiveness domains. If the purpose of the analysis were to find whether or not the original survey variables could be reduced to a small number of latent constructs, then we would use Exploratory Factor Analysis (EFA). EFA is used when there is no a priori assumption about the underlying dimensionality of the construct. CFA has stronger requirements than exploratory factor analysis. In particular, one must specify the number of hypothesized factors and say in advance which items belong on which factors. Perhaps more important is that CFA does not try to avoid dealing with measurement error, but instead considers it in developing the factor loadings (Hurley, A. E. et al., 1997). Different methods of factor analysis have their strengths and weaknesses, but the choice of which to use of one depends on the requirements and context of the study.

Ones again, we want to confirm theoretical assumption that a set of questions measuring people's experience with the health systems can be expressed in a small number of responsiveness domains. We

assume there are seven domains of responsiveness in relation to experience of ambulatory care services. First, we will assess whether the items in the responsiveness module can be summarised into seven responsiveness domains, next, we will assess whether all the items can be summarised in a single latent construct that we will define as overall ambulatory care responsiveness. Each of the seven responsiveness domains for ambulatory care experience (defined as Prompt attention, Dignity, Communication, Autonomy, Confidentiality, Choice and Quality of basic amenities) was treated as a separate construct. Only those questions hypothesised to relate to a particular domain were included in the model with requirement to attain only 1 single factor.

We will assess the construct validity by factor analysis the principal factor method used to analyse the correlation matrix. The factor loadings, sometimes called the factor patterns are computed using the squared multiple correlations as estimates of the communality. There is no strict cut off to describe strong and weak associations of variance, but the closer to +1 or -1, the stronger the unidimensionality of the construct is. For example Gasquet (2004) and Haddad (1998, 2000) set in their studies as the substantial factor loading ≥ 0,4, for Westaway (2003) loading >0,5. Table 1 summarizes results for all seven domains monitored in ambulatory care. Factors names were given according to each responsiveness domain. The results show all items have very high correlation with the responsiveness domains. Only q6103 loaded slightly lower, what we can assign to different scale than in all other evaluated items. Nevertheless, this question still has a reasonable correlation of 0,38 with the prompt attention. All seven factors (or responsiveness domains) achieved eigenvalues greater than 1. The results indicate that the construct we have used is a valid one.

Table 1: **Factor Loadings for ambulatory care**

Item	Responsiveness Domain	Factor loading
q6101	Prompt attention	0.59
q6103	Prompt attention	0.38
q6104	Prompt attention	0.68
q6110	Dignity	0.87
q6111	Dignity	0.83
q6112	Dignity	0.74
q6113	Dignity	0.78
q6120	Communication	0.80
q6121	Communication	0.81
q6122	Communication	0.83
q6123	Communication	0.82
q6131	Autonomy	0.76
q6132	Autonomy	0.74
q6133	Autonomy	0.80
q6140	Confidentiality	0.72
q6141	Confidentiality	0.80
q6142	Confidentiality	0.84
q6150	Choice	0.87
q6151	Choice	0.89
q6152	Choice	0.87
q6160	Quality of basic amenities	0.74
q6161	Quality of basic amenities	0.79
q6162	Quality of basic amenities	0.82

Unidimensionality, or internal consistency of items, can be also assessed by using Cronbach's alpha. Cronbach's alpha measures how well a set of items measures a single unidimensional latent construct, the coefficient measures overall correlation between the items and the scale. Cronbach's alpha is more a coefficient of reliability than a statistical test that means the higher the coefficient, the higher is the unidimensionality of the items. The alpha coefficient ranges from 0 (lowest reliability) to 1 (highest reliability) and a value of 0,8 or greater is considered good, meaning that the items measure a single unidimensional construct (Sitzia, J., 1999). According to studies of Hagen (2003), Blazeby (2004), Labarere (2001), Verho (2003), Jenkinson (2002, 2003), Li (2003) or Snelgrove and Slater (2003), it is sufficient to reach the criterion alpha value 0,7 to be confident about unidimensionality. The lowest criterion for alpha was set by Steine (2001) to the value of 0,6. On the other hand, Salomon (1999) set the criterion value to 0,8.

Internal consistency using Cronbach`s alpha was also tested by authors evaluating sociological surveys (Snelgrove, S. – Slater, J., 2003; Johnson, L. M. et al., 2002; Speer, P. W. – Peterson, N. A., 2000; Řehák, J., 1998a; Řehák, J., et al., 1998; Gentry, M. – Owen, S. V., 2004). Cronbach`s alpha test also offers results for inter – item correlation and authors as Boling (2003), Hagen (2003), Blazeby (2004), McGuiness (2003), Paisley (2001), Westaway (2003), Gasquet (2004), Jenkinson (2002, 2003), Li (2003) or Snelgrove and Slater (2003) suggested the inter - item correlation coefficient should be higher than 0,4. From the authors evaluating the sociological surveys the inter – item correlation was tested by Choi (2004), Snelgrove and Slater (2003), Johnson (2002), Speer and Peterson (2000), Řehák (1998a, 1998b), Řehák et al. (1998) and Gentry and Owen (2004). Table 2 shows the results for inter – item correlations and alpha values for individual domains for ambulatory care patients. Nevertheless you can observe a moderate mean inter – item correlation for the prompt attention domain (0,3383); the alpha coefficient is still quite high 0,6054. The rest of the alpha coefficients reached value higher than 0,8. The results again indicated that the construct we have used is a valid one.

Table 2: **Cronbach's alpha coefficients and inter – item correlation for ambulatory care**

Responsiveness Domain	**Inter - item correlation**	**Alpha**
Prompt attention	0.3383	0.6054
Dignity	0.6657	0.8884
Communication	0.6909	0.8994
Autonomy	0.6335	0.8384
Confidentiality	0.6558	0.8511
Choice	0.7952	0.9209
Quality of basic amenities	0.6537	0.8499

Next, we explore if data of the ambulatory care responsiveness items are a measure of a single construct. We actually want to find a meaningful construct that is a surrogate for ambulatory care responsiveness from the 23 questions presented in Table 1. We will also assess whether all these items have unidimensionality with the help of Cronbach's alpha coefficient.

Then we look at the factor loadings (Table 3), the standardised correlation coefficients between responses to the questions and the construct or the factor gained through the factor analysis the principal factor method with requirement to attain just 1 single factor, we see that except for 2 questions items correlate highly (0,5 or higher) to the factor. Although result for q6103 shows quite poor correlation with the latent construct (factor explains only 12% of the variance in this variable), the other questions highly correlate with the overall ambulatory care responsiveness. The factor (or overall responsiveness) achieved eigenvalues greater than 1 and accounted for 69,5% of the total variance. Thus we can conclude that these 23 items can be expressed in a single scale, which we refer to ambulatory care responsiveness.

Table 3: **Factor Loadings for ambulatory care – single construct**

Variable	*Factor Loadings 1*	*Uniqueness*
q6101	*0.56782*	*0.67758*
q6103	*0.34310*	*0.88228*
q6104	*0.71717*	*0.48567*
q6110	*0.74867*	*0.43949*
q6111	*0.69482*	*0.51722*
q6112	*0.67891*	*0.53908*
q6113	*0.72414*	*0.47563*
q6120	*0.80478*	*0.35234*
q6121	*0.70797*	*0.49877*
q6122	*0.73976*	*0.45276*
q6123	*0.79683*	*0.36505*
q6131	*0.66528*	*0.55740*
q6132	*0.61629*	*0.62018*
q6133	*0.79436*	*0.36899*
q6140	*0.56235*	*0.68377*
q6141	*0.73136*	*0.46512*
q6142	*0.74851*	*0.43973*
q6150	*0.70827*	*0.49835*
q6151	*0.66739*	*0.55459*

q6152 |	*0.76025*	*0.42202*
q6160 |	*0.48322*	*0.76649*
q6161 |	*0.54079*	*0.70754*
q6162 |	*0.58206*	*0.66120*

Cronbach's alpha coefficient suggests whether we can summarize these 23 items into a single factor. The reliability coefficient measured by Cronbach's alpha shows how well these items measure a single unidimensional latent construct. The scale reliability coefficient is very high, at 0,94, suggesting the unidimensional nature of the data. In other words, the data from 23 items can be expressed as unidimensional, which we termed as overall outpatient responsiveness (Table 4).

Table 4: **Cronbach`s alpha**

Average inter item covariance:	*0.3125*
Number of items in the scale:	*23*
Scale reliability coefficient:	*0.9392*

Reliability is a function of random measurement error that is unpredictable and occurs in all surveys. Reliability testing uses data from a population to estimate the portion of the variance that is true or non-random; this proportion is expressed as a coefficient between 0 and 1. Reliability is a statistical measure that tells us how reproducible the data from a survey instrument is. Reliability is concerned with the consistency of the measurement, the degree to which the questionnaire measures the same way each time it is used under the same conditions with the same respondents. There are many forms of reliability, such as test-retest reliability, alternate-form reliability, internal consistency reliability, inter-observer reliability or intra-observer reliability (Litwin, M. S., 1995). Test-retest reliability is most of the researchers interest, also to sociological (Blazeby, J. M. et al., 2004; McGuiness, C. – Sibthorpe, C., 2003; Li, L. et al., 2003; Gigantesco, A. et al., 2003; Hendriks, J. et al., 2002; Salomon, L. et al., 1999; Gasquet, I. et al., 2004; Terry, R., 2000; Řehák, J., 1998a; Řehák, J., 1998b; Řehák, J. et al., 1998) and estimates the error component when there is repetition of a measurement by computing Kappa statistics for categorical and intra-class correlation coefficient for continuous variables within and across populations. This gives us estimates of agreement rates for concordance between test and retest applications to indicate the stability of the application. That means, for example, if the same question is asked to the same respondent at two different times and produces the same results, then the Kappa statistic will have a value of 1. A score of 1 indicates perfect concordance between the two sets of responses; a score of 0 indicates that the observed concordance was no better than could be expected by chance. Landis and Koch (1977) provide the next guideline for interpreting ranges of Kappa values, below 0 means poor, between 0 – 0,20 slight, 0,21 – 0, 40 fair, 0.41 – 0,60 moderate, 0,61 – 0,80 substantial and between 0,81 – 1 as almost perfect reliability of tested questionnaire. For example, Salomon (1999) and Gasquet (2004) set in their studies the satisfactory value of the Kappa coefficient equal to 0,6.

Because WHO do not offer pre-pilot dataset we will reproduce results presented in Murray and Evans (2003). Slovak respondents were approached one week after the first questionnaire was administered. A total of 96 retest interviews were performed. Of these 74 individuals reported having had ambulatory care experiences in the previous 12 months, 2 had home care, 16 had hospital inpatient care, and the remainder had no care experiences. Data where the retest involved less than 30 interviews were omitted from analysis. Analysing the given result in Table 5 we can conclude, that the reliability of the instrument in ambulatory care domains was substantial or almost perfect (Landis, J. R. – Koch, G. G., 1977). The average Kappa for Slovakia was 0,81; that means the level of agreement between the responses to the original survey and the retest interviews was almost perfect. We can see that the ambulatory care domains attained Kappa values higher than 0,8. We can conclude that the questionnaire had high reproducibility.

Table 5: **Kappa rates for ambulatory care calculated from retests**

Responsiveness Domain	**Kappa**
Prompt attention	0.87
Dignity	0.82
Communication	0.82
Autonomy	0.90

Confidentiality	0.91
Choice	0.91
Quality of basic amenities	0.88
Average Kappa per Slovakia	0.81

Source: Murray CJL, Evans DB (Eds) Health systems performance assessment: debates, methods and empiricism. Geneva (2003: p.606)

For a survey to be feasible, it has to actually work in the field. To measure it, we look at the response rates and the missing values of the MCSS questionnaire.

Response rates should be maximised, as incomplete responses contribute to uncertainty about the generalisability of the findings from the survey sample to the population from which the survey is drawn. Sitzia & Wood (1998) in their article conclude that there is no basis for establishing an 'acceptable' response rate. However, the practise showed different survey types bring different response rates. Higher response rate is usually gained when face-to-face or phone questionnaire is administered, lower response rate occurs when researchers use postal or drop – off types of surveys (Hagen, S. et al., 2003; Blazeby, J. M. et al., 2004; McGuiness, C. – Sibthorpe, C., 2003; Labarere, J. et al., 2001; Li, L. et al., 2003, Jenkinson, C. et al., 2003; Jenkinson, C. et al., 2002; Haddad, S. et al., 2000; Westaway, M. S. et al., 2003; Gigantesco, A. et al., 2003; Paddock, L. E. et al., 2000; Hendriks, J. et al., 2002; Salomon, L. et al., 1999; Steine, S. et al., 2001; Blazeby, J. M. et al., 2004). Snelgrove and Slater (2003), Terry (2000), Gentry and Owen (2004) and Speer and Peterson (2000) indicated in their surveys achieved response rate. We will again reproduce the results published by Murray and Evans (2003). The WHO MCSS in Slovakia was administered as face-to-face household survey reaching 84% response rate on average. We can conclude that high response rate supports questionnaire feasibility.

There are a number of reasons why data are missing from questionnaires, such as skip patterns may not be properly followed, filter questions may be inadequately administered, problems may arise with question wording or inaccurate data inputting. The assessment of this kind of problem is known as item missing analysis. Labarere (2001) suggested that the proportion of missing values should not be higher than 20%. WHO has pre-established a cut-off point of 20% for missing rates, meaning that any question that has a missing rate of 20% or more, becomes a problematic question. Whatever the reasons, if there is a missing rate 20% or greater for any question, it should not be used for further analysis without a thorough technical explanation. A missing rate is in general defined as the percentage of non-responses to an item, refusals to answer, and responses of "not applicable" and "don't know". Responsiveness module ambulatory care did not have the possibility to answer "not applicable" or "don't know", so if respondent did not know the answer or it was irrelevant, he/she just did not answer and the answer was recorder as missing value. We have not found in reviewed sociological literature any author indicating the level of missing rates.

When we advance in missing analysis we have to bear in mind skips, otherwise high missing rates appear that are actually incorrect. For example, in the MCSS responsiveness module we find that there are two questions mentioned earlier in the questionnaire that are filters. Respondents are only required to answer and rate their experience if they received any ambulatory health care in the last 12 months. There are questions (q6000 and q6001), which may direct the respondents to skip questions about experiences with health care. That means if a respondent answers 'No' to these questions, he/she skips all the questions in the section related to ambulatory care health system. Table 6 shows missing rates for ambulatory care items reflecting already mentioned filters. Four questions exceeded allowed 20% cut-off and these questions need proper reasoning before involving into questionnaire evaluation. Except for these four questions the responsiveness module ambulatory care met the pre-set criteria for a missing rate of less than 20%. The four questions not meeting these criteria should be amenable to correction in future survey rounds.

Table 6: **Missing rates for ambulatory care**

Item	Responsiveness Domain	Missing rate in %
q6101	Prompt attention	3.27
q6103	Prompt attention	38.57
q6104	Prompt attention	1.51
q6110	Dignity	0.63
q6111	Dignity	5.91
q6112	Dignity	1.51
q6113	Dignity	0.50

q6120	Communication	0.13
q6121	Communication	1.01
q6122	Communication	0.75
q6123	Communication	0.38
q6131	Autonomy	46.36
q6132	Autonomy	46.73
q6133	Autonomy	3.64
q6140	Confidentiality	2.39
q6141	Confidentiality	4.90
q6142	Confidentiality	3.14
q6150	Choice	12.06
q6151	Choice	32.99
q6152	Choice	11.56
q6160	Quality of basic amenities	0.38
q6161	Quality of basic amenities	0.13
q6162	Quality of basic amenities	0.50

Conclusion

We presented the psychometric properties or methodological requirements for data collected through the questionnaire surveys that should be satisfied prior to any statistical manipulation, evaluation and final interpretation of achieved results gained through the questionnaire surveys. According to Řehák (1998b), the quality of data determinate the quality of results. As an example of presented methodology, we showed the psychometric properties of the MCSS Responsiveness module – ambulatory care Slovak dataset collected by World Health Organisation in 2001. We evaluated the quality of observed data, in concrete the validity, reliability and feasibility of the survey questionnaire. Just to notice, all questionnaires, does not matter if they survey economics, health or sociological area must satisfy psychometric properties. Validity was tested by factor analysis the principal factor method to find out whether or not the survey data was unidimensional. The factor analysis results generally confirmed not only the assumed structure of the responsiveness domains, but also the overall ambulatory care responsiveness. The validity of the scale where a number of questions were asked to measure each of the responsiveness domains or one construct was supported by data. The factor loadings for domains, except for two items, were greater than 0,6. We suggest reviewing those two questions (q6101 and q6103) in next survey round. In case of one single construct only two items reached factor loading lower than 0,5. Also in this case we suggest changing the questions (q6160 and q6103, change in wording and scale). Unidimensionality of the data was proved also by using Cronbach's Alpha. Only prompt attention domain reached slowly lower alpha coefficient (less than 0,7) and inter – item correlation (less than 0,4). The result is understandable and we have already mentioned demand to review questions q6101 and q6103. Cronbach`s alpha as well as inter – item correlations again suggest to review those two questions. Except those few questions the analyses confirmed that the items measure a single unidimensional construct. Reliability is concerned with the consistency of the measurement, the degree to which an instrument measures the same way each time it is used under the same conditions with the same respondents. Test-retest results proved the necessary quality (Kappa higher than 0,8) and we conclude the reliability of the questionnaire in ambulatory care domains was almost perfect. Feasibility is connected to results obtained in the field. We evaluate the reached response rates and the missing values of each item of the questionnaire. The MCSS attained 84% response rate concluding a very good survey result. In case of missing rates four items in responsiveness module ambulatory care exceeded defined 20% cut-off suggesting giving special attention to these items prior involving them into the questionnaire evaluation. In case we are not able to reasonable explain higher than 20% missing rate, we have to reject those questions from further statistical evaluation (and also is advisable to modify those questions in next survey round).

Appendix 1

Exact question wording of the MCSS Responsiveness Ambulatory Care Questions

Domain	Question number	Question wording	Type of response scale
Prompt attention	Q6101	In the last 12 months, when you wanted care, how often did you get care as soon as you wanted?	Frequency Reporting (Always - Never)
	Q6103	Generally, how long did you have to wait before you could get the laboratory tests or examinations done?	Other Reporting (number of days)
	Q6104	Now, overall, how would you rate your experience of getting prompt attention at the health services in the last 12 months?	Rating (Very Good – Very Bad)
Dignity	Q6110	In the last 12 months, when you sought health care, how often did doctors, nurses or other health care providers treat you with respect?	Frequency Reporting (Always – Never)
	Q6111	In the last 12 months, how often did the office staff, such as receptionists or clerks there, treat you with respect?	Frequency Reporting (Always – Never)
	Q6112	In the last 12 months, how often were your physical examinations and treatments done in a way that your privacy was respected?	Frequency Reporting (Always – Never)
	Q6113	Now, overall, how would you rate your experience of being treated with dignity at the health services in the last 12 months?	Rating (Very Good – Very Bad)
Clear communication	Q6120	In the last 12 months, how often did doctors, nurses or other health care providers listen carefully to you?	Frequency Reporting (Always – Never)
	Q6121	In the last 12 months, how often did doctors, nurses or other health care providers, explain things in a way you could understand?	Frequency Reporting (Always – Never)
	Q6122	In the last 12 months, how often did doctors, nurses, or other health care providers give you time to ask questions about your health problem or treatment?	Frequency Reporting (Always – Never)
	Q6123	Now, overall, how would you rate your experience of how well health care providers communicated with you in the last 12 months?	Rating (Very Good – Very Bad)
Autonomy	Q6131	In the last 12 months, how often did doctors, nurses or other health care providers involve you as much as you wanted in deciding about the care, treatment or tests?	Frequency Reporting (Always – Never ever)
	Q6132	In the last 12 months, how often did doctors, nurses or other health care providers ask your permission before starting the treatment or tests?	Frequency Reporting (Always – Never)
	Q6133	Now, overall, how would you rate your experience of getting involved in making decisions about your care or treatment as much as you wanted in the last 12 months?	Rating (Very Good – Very Bad)
Confidentiality	Q6140	In the last 12 months, how often were talks with your doctor, nurse or other health care provider done privately so other people who you did not want to hear could not overhear what was said?	Frequency Reporting (Always – Never)
	Q6141	In the last 12 months, how often did your doctor, nurse or other health care provider keep your personal information confidential? This means that anyone whom you did not want informed could not find out about your medical conditions.	Frequency Reporting (Always – Never)
	Q6142	Now, overall, how would you rate your experience of the way the health services kept information about you confidential in the last 12 months?	Rating (Very Good – Very Bad)
Choice of health care provider	Q6150	Over the last 12 months, with the doctors, nurses and other health care providers available to you how big a problem, if any, was it to get a health care provider you were happy with?	Other Reporting (Level of problem)
	Q6151	Over the last 12 months, how big a problem, if any, was it to get to use other health services other than the one you usually went to?	Other Reporting (Level of problem)
	Q6152	Now, overall, how would you rate your experience of being able to use a health care provider or service of your choice over the last 12 months?	Rating (Very Good – Very Bad)

Domain	Question number	Question wording	Type of response scale
Quality basic amenities	Q6160	Thinking about the places you visited for health care in the last 12 months, how would you rate the basic quality of the waiting room, for example, space, seating and fresh air?	Rating (Very Good – Very Bad)
	Q6161	Thinking about the places you visited for health care over the last 12 months, how would you rate the cleanliness of the place?	Rating (Very Good – Very Bad)
	Q6162	Now, overall, how would you rate the quality of the surroundings, for example, space, seating, fresh air and cleanliness of the health services you visited in the last 12 months?	Rating (Very Good – Very Bad)

Response scale options

Report (Frequency) scales: Always, Usually, Sometimes, Never
Rating scales: Very good, Good, Moderate, Bad, Very bad
Other reporting (number of days): Same day, 1-2 days, 3-5 days, 6-10 days, More than 10 days
Other rating (problems): No problem, Mild, Moderate, Severe, Extreme Problem

Source: MCSS Long 90 min Questionnaire (www.who.int)

Hana Zach *graduated in Economics and is PhD student in Economics at Department of Statistics and Operation Research, Slovak University of Agriculture in Nitra (Slovak Republic) as well as in Biomedical Informatics at 1st Faculty of Medicine, Charles University in Prague (Czech Republic). Her research focuses on applied statistics in economics, environmental issues and health related topics.*

BIBLIOGRAPHY

BLAND, J. M. – ALTMAN, D. G., 1997: Statistics notes: Cronbach's alpha. In: British Medical Journal, 314(7080), pp. 572-572.

BLAND, J. M. – ALTMAN, D. G., 2002: Statistics Notes: Validating scales and indexes. In: British Medical Journal, 324(7337), pp. 606-607.

BLAZEBY, J. M. – CONROY, T. – BOTTOMLEY, T. – VICKERY, C. – ARRARAS, J. – SEZER, O. – MOORE, J. – KOLLER, M. – TURHAL, N. S. – STUART, R. –VAN CUTSEM, E. – D'HAESE, S. – COENS, C., 2004: On behalf of the European Organisation for Research and Treatment of Cancer Gastrointestinal and Quality of Life Groups Clinical and psychometric validation of a questionnaire module, the EORTC QLQ-STO 22, to assess quality of life in patients with gastric cancer. In: European Journal of Cancer, 40(15), pp. 2260-2268.

BOLING, W. – FOULADI, R. T. – BASEN-ENGQUIS, K., 2003: Health-related quality of life in gynecological oncology: instruments and psychometric properties. In: International Journal of Gynecological Cancer, 13(1), pp. 5-14.

DE VELLIS, R., 1991: Scale Development: Theory and Applications. Newbury Park, Sage.

CHOI, N., 2004: A Psychometric Examination of the Personal Attributes Questionnaire. In: Journal of Social Psychology, Vol. 144, Issue 3, pp. 348-352.

GASQUET, I. – VILLEMINOT, S. – ESTAQUIO, C. – DURIEUX, P. – RAVAUD, P. – FALISSARD, P., 2004: Construction of a questionnaire measuring outpatients' opinion of quality of hospital consultation departments. In: Health and Quality of Life Outcomes. 2(1), p. 43.

GENTRY, M. – OWEN, S. V., 2004: Secondary Student Perceptions of Classroom Quality: Instrumentation and Differences Between Advanced/Honors and Nonhonors Classes. In: Journal of Secondary Gifted Education, Vol. 16, Issue 1, pp. 20-29.

GIGANTESCO, A. – MOROSINI, P. – BAZZONI, A., 2003: Quality of psychiatric care: validation of an instrument for measuring inpatient opinion. In: International Journal for Quality in Health Care, 15, pp. 73-78.

HADDAD, S. – FOURNIER, P. – POTVIN, L., 1998: Measuring lay people's perceptions of the quality of primary health care services in developing countries. Validation of a 20-item scale. In: International Journal for Quality in Health Care, 10, pp. 93-104.

HADDAD, S. – POTVIN, L. – ROBERGE, D. – PINEAULT, R. – REMONDIN, M., 2000: Patient perception of quality following a visit to a doctor in a primary care unit. In: Family Practice, 17, pp. 21-29.

HAGEN, S. – BUGGE, C. – ALEXANDER, H., 2003: Psychometric properties of the SF-36 in the early post-stroke phase. In: Journal of Advanced Nursing, 44(5), pp. 461-468.

HENDRIKS, J. – OORT, F. J. – VRIELINK, M. R. – SMETS, E. M. A., 2002: Reliability and validity of the Satisfaction with Hospital Care Questionnaire. In: International Journal for Quality in Health Care, 14, pp. 471-482.

HURLEY, A. E. – SCANDURA, T. – SCHRIESHEIM, C. A. – BRANNICK, M. – SEERS, A. – VANDENBURG, R. – WILLIAMS, L. J., 1997: Exploratory and confirmatory factor analysis: guidelines, issues and alternatives. In: Journal of Organizational Behaviour, 18, pp. 667-683.

JENKINSON, C. – COULTER, A. – BRUSTER, S., 2002: The Picker Patient Experience Questionnaire: development and validation using data from in-patient surveys in five countries. In: International Journal for Quality in Health Care, 14, pp. 353-358.

JENKINSON, C. – FITZPATRICK, R. – NORQUIST, J. – FINDLEY, L. – HUGHES, J., 2003: Cross-cultural evaluation of the Parkinson's Disease Questionnaire: tests of data quality, score reliability, response rate, and scaling assumptions in the United States, Canada, Japan, Italy, and Spain. In: Journal of Clinical Epidemiology, 56(9), pp. 843-847.

JOHNSON, L. M. – MULLICK, R. – MULFORD, CH. L., 2002: General versus specific victim blaming. In: Journal of Social Psychology, Vol. 142 (2), pp. 249-263.

LABARERE, J. – FRANCOIS, P. – AUQUIER, P. – ROBERT, C., 2001: Development of a French inpatient satisfaction questionnaire. In: International Journal for Quality in Health Care, 13, pp. 99-108.

LANDIS, J. R. – KOCH, G. G., 1977: The measurement of observer agreement for categorical data. In: Biometrics, 33(1), pp. 159-174.

LI, L. – WANG, H. M. – SHEN, Y., 2003: Chinese SF-36 Health Survey: translation, cultural adaptation, validation, and normalisation. In: Journal of Epidemiology & Community Health, 57(4), pp. 259-263.

LITWIN, M. S., 1995: How to measure survey reliability and validity (The Survey Kit series). Thousand Oaks, CA, Sage.

MCGUINESS, C. – SIBTHORPE, C., 2003:. Development and initial validation of a measure of coordination of health care. In: International Journal for Quality in Health Care, 15, pp. 309-318.

MURRAY, C. J. L. – EVANS, D. B. (Eds), 2003: Health systems performance assessment: debates, methods and empiricism. Geneva, WHO. ISBN 92 4156245 5.

PADDOCK, L. E. – VELOSKI, J. – CHATTERTON, M. L. – GEVIRTZ, F. O. – NASH, D. B., 2000: Development and validation of a questionnaire to evaluate patient satisfaction with diabetes disease management. In: Diabetes Care, 23, pp. 951-956.

PAISLEY, A. M. – BALDWIN, P. – PATERSON-BROWN, P., 2001: Feasibility, reliability and validity of a new assessment form for use with basic surgical trainees. In: The American Journal of Surgery, 182(1), pp. 24-29.

ŘEHÁK, J., 1998a: Kvalita dat I. In: Sociologický časopis, Vol. 34, No. 1, pp. 51-60.

ŘEHÁK, J., 1998b: Kvalita dat II. In: Sociologický časopis, Vol. 34, No. 2, pp. 195-204.

ŘEHÁK, J. – BÁRTOVÁ, I. – HAMANOVÁ, J., 1998: Kvalita dat III. In: Sociologický časopis, Vol. 34, No. 3, pp. 363-372.

ŘEHÁK, J. – ŘEHÁKOVÁ, B., 1986: Analýza kategorizovaných dat v sociologii. Praha, Academia.

SALOMON, L. – GASQUET, I. – MESBAH, M. – RAVAUD, P., 1999: Construction of a scale measuring inpatients' opinion on quality of care. In: International Journal for Quality in Health Care, 11, pp. 507-516.

SITZIA, J., 1999: How valid and reliable are patient satisfaction data? An analysis of 195 studies. In: International Journal for Quality in Health Care, Vol. 11, No. 4, pp. 319-328.

SITZIA, J. – WOOD, N., 1998: Response rate in patient satisfaction research: an analysis of 210 published studies. In: International Journal for Quality in Health Care, Vol. 10, No. 4, pp. 311-317.

SNELGROVE, S. – SLATER, J., 2003: Approaches to learning: psychometric testing of a study process questionnaire. In: Journal of Advanced Nursing, Vol. 43, Issue 5, p 496-505.

SPEER, P. W. – PETERSON, N. A., 2000: Psychometric properties of an empowerment scale: Testing cognitive, emotional, behavioural domains. In: Social Work Research, Vol. 24, Issue 2, pp. 109-118.

STEINE, S. – FINSET, A. – LAERUM, E., 2001: A new, brief questionnaire (PEQ) developed in primary health care for measuring patients' experience of interaction, emotion and consultation outcome. In: Family Practice 18(4), pp. 410-418.

SURYNEK, A. – KOMÁRKOVÁ, R. – KAŠPAROVÁ, E., 2001: Základy sociologického výskumu. Praha, Managemnt Press. ISBN 80-7261-038-4.

TERRY, R., 2000: Recent Advances in Measurement Theory and the Use of Sociometric Techniques. In: New Directions for Child & Adolescent Development, Summer (88), pp.27-53.

VERHO, H. – ARNETZ, J. E., 2003: Validation and application of an instrument for measuring patient relatives' perception of quality of geriatric care. In: International Journal for Quality in Health Care, 15, pp. 197-206.

WESTAWAY, M. S. – RHEEDER, P. – VAN ZYL, D. G. – SEAGER, J. R., 2003: Interpersonal and organizational dimensions of patient satisfaction: the moderating effects of health status. In: International Journal for Quality in Health Care, 15, pp. 337-344.

World Health Organisation. (www.who.int) Retrieved 5-2-2005.