How Does a Change in the Administration Method Affect the Reliability of the COOP /WONCA Charts?

Cindy L K Lam, FRCGP, FHKAM (Family Medicine), Associate Professor[1]

Ian J Lauder , Ph.D (Statistics), Associate Professor[2]

Daniel T P Lam, FRACGP, FHKAM (Family Medicine), Associate Professor[1]

The University of Hong Kong

1. General Practice Unit, 3rd Floor, Ap Lei Chau Clinic, 161 Main Street, Ap Lei Chau, Hong Kong.
Fax: (852) 28147475, Tel: (852) 25526021, E mail: *clklam@hku.hk*

2. Department of Statistics, the University of Hong Kong, Pokfulam, Hong Kong

How Does a Change in the Administration Method Affect the Reliability of the COOP/WONCA Charts?

Abstract

Background: An interviewer is often needed to administer the COOP/WONCA Charts to Chinese patients and this may affect the reliability of results. **Objectives:** To find out the reliability of the COOP/WONCA Charts administered by an interviewer, and whether a change in the interviewer or administration method would affect the results. **Methods**: A cross sectional test-retest study on 487 Chinese adult patients attending a family medicine clinic in Hong Kong. The COOP/WONCA Charts were administered either by the same interviewer, two different interviewers, or self-completion and interviewer administration, on test and retest. The random, inter-observer and inter-method variances were compared to the inter-subject variance. The reliability coefficient of each COOP/WONCA Chart was calculated for each method of administration. **Results**: Random errors could change the scores by 0.57 to 1.04, inter-observer variations could change the scores of four Charts by 0.72 to 0.80, and a change in the method could change the physical fitness score by 1.79 and the daily activities score by 1.31, on a 5 point scale. The reliability coefficients of the six COOP/WONCA Charts were 0.68 to 0.92 for one interviewer, 0.59 to 0.82 for two interviewers, and 0.46 to 0.81 for two methods. **Conclusion**: The Chinese COOP/WONCA Charts were reliable in detecting real differences when administered by an interviewer. A change in the method of administration significantly decreased the reliability of the results. The use of more than one method of data collection in the same survey should be discouraged.

Keywords: Reliability, COOP/WONCA Charts, Functional health, Chinese.

Introduction

The Dartmouth COOP functional health assessment Charts/WONCA (COOP/WONCA Charts) are a popular instrument for the measurement of functional

status in primary care [1-3]. They were first developed by Nelson et al and later modified by the Classification Committee of the World Organization of Family Doctors (WONCA)[1-3]. There are six charts, one each on physical fitness, feelings, daily activities, social activities, change in health and overall health. Each chart is rated on a five-point scale with higher scores indicating worse functional status. The COOP/WONCA Charts have been translated and validated for many cultures including the Chinese[3,4]. It is commonly used for comparing the health status between patient groups, monitoring changes in functional status over time, and measuring the outcomes of interventions.

The COOP/WONCA Charts can be administered by self or an interviewer. Scholten et al proposed self-completion to be the method of choice to avoid observer (interviewer) bias[2]. However, this method is not feasible for people who are illiterate. Thirteen percent of the general population and 43% of those aged 55 years or over in Hong Kong are illiterate, the rates are even higher in mainland China[5,6]. The Charts often need to be administered by an interviewer when they are applied to these Chinese populations and this raises a concern for the reliability of results. Nelson et al showed that the original Dartmouth COOP Charts had good one-hour test-retest reliability when administered by one or more interviewers to American patients[1], but this has not been tested on the revised COOP/WONCA Charts and the technical equivalence of self-completion and interviewer administration has never been assessed.

The aim of our study was to find out if the COOP/WONCA Charts were reliable when administered to Chinese subjects by an interviewer. We also wanted to find out how a change in the interviewer or interviewing method would affect the scores.

Ideally, the same result should be obtained on repeated assessments of the same individual in the same situation irrespective of the observer or measurement method. Unfortunately, variations in measurements are inevitable even if they were done by the same observer and method due to random and replicative errors[7,8] . The subjective nature of health status assessment makes it more liable to variations because people's perception may change with time and the environment. Different interviewers may lead to different responses because their attitudes, communication skills and personal preferences may influence a subject's perception. The interpretation of the questions and response choices could be different when they are administered by self or an interviewer, leading to different results.

An observed difference or change over time could be the result of measurement variation[7,8]. This has great implication when health assessment is used as an evaluative or outcome measure. We need to know the magnitude of the measurement errors before we can decide whether an observed difference is significant or not. An assessment instrument is reliable if any difference detected is predominantly due to a true difference between subjects or a real change over time. It is useless if measurement errors are greater than true differences.

Subjects and Methods

The study was carried out in a family medicine clinic that had two full-time and two part-time doctors serving a population of 5000 Chinese people in Hong Kong. Data collection was carried out in three phases, all adult patients (aged 18 years or over) attending the clinic during the specified survey periods were invited to take part, each patient could be included in only one phase of the study. Table 1 shows the characteristics of the patient samples in the three phases of the study. We used a test-retest study design in that each subject answered the Chinese version of the COOP/WONCA Charts[4] before and after his/her doctor consultation.

The first phase (two-interviewer) was designed for the assessment of the inter-observer variance (Vo). 84 patients were randomly assigned to be interviewed by the same (n=40), or two different (n=44), interviewers on test and retest. The inter-observer variance (Vo) was estimated from the paired test-retest score variance of the two-interviewer group after controlling for the variance of the same-interviewer group.

The second phase surveyed 195 patients who said that they could read and write. They completed the charts first by self-completion and then the charts were administered by an interviewer (two-method sample). A change from self-completion to interviewer-administration involved a change in the observer as well as a change in the method. The inter-method variance (Vm) was estimated from the paired test-retest score variance of the two-method sample after controlling for the two interviewer variance found in the first phase of the study.

The third phase surveyed 208 patients with the COOP/WONCA Charts administered by the same interviewer in both test and retest (one-interviewer sample). The data were used to assess the intra-observer random replicative variance (Vr), and

the inter-subject variance (V). Vr was calculated from the differences between the paired test-retest scores, and V was obtained by excluding Vr from the total variance.

The standard technique of analysis of variance (ANOVA) was used to determine the variance components by equating the computed mean squares with their expected values from ANOVA theory[9]. The standard F test for variance ratios was used to compare the different variance components at the 5% level of significance. Since variance is the square of standard deviation, the 95% confidence interval of the score change was estimated to be ± 2 times the squared-root of the variance.

We calculated the reliability coefficients of each COOP/WONCA Chart by dividing the true (inter-subject) variance by the total variance for one interviewer, two interviewers and two methods, respectively [7,8]. The reliability coefficient is a measure of the reliability of the instrument in detecting true differences. The most widely accepted standard is 0.7 or more for group comparison[10], although Helmstadter has proposed a lower standard of 0.5[8].

The Wilcoxon matched-pairs signed-ranks test of the SPSS for Windows programme was used to test if there was any significant bias in the retest scores.

Results

Table 2 shows the inter-subject variance (V), intra-observer random replicative variance (Vr), inter-observer variance (Vo) and inter-method variance (Vm), and their corresponding 95% confidence intervals of score changes, for the six COOP/WONCA Charts. All the COOP/WONCA Charts were scored on a five-point scale. Random replicative errors could cause changes in the chart scores of 0.57 to 1.04. A change in the observer could cause additional changes of 0.72 to 0.80 in the scores of the physical fitness, daily activities, social activities and overall health charts. The random and observer variations together could change the scores up to 1.81 (daily activities chart) when there was a change in the interviewer. A change in the method of administration could further change the physical fitness score by 1.79 and the daily activities score by 1.31. The total measurement variations could change the physical fitness and daily activities scores by more than 3 when the administration method was changed.

Table 3 shows the reliability coefficients of the COOP/WONCA Charts for the same interviewer, two interviewers and two methods, respectively. Five charts had coefficients greater than 0.7 and only one (change in health) chart was marginally below the standard when the charts were administered by the same interviewer. The reliability coefficients of three charts were below 0.7 but all were above 0.5 when they were administered by two interviewers. When two methods were used, the reliability coefficients of only two charts were above 0.7, three were between 0.5 to 0.7, and that of the daily activities chart was less than 0.5.

Table 4 shows the paired differences in the test-retest scores of the COOP/WONCA Charts when they were administered by the same interviewer, two

different interviewers or two different methods. The test-retest concordance (no change in score) rates were all above 75% with few score changes of more than one when the COOP/WONCA Charts were administered by the same interviewer. There was a tendency for the retest scores to be better than the test scores for the feelings and daily activities charts when they were administered by the same interviewer. The two-interviewer concordance rates of most Charts were lower (59%-86%) than those achieved by the same interviewer, but there was no significant bias in the retest scores. The concordance rates between the scores of self-completion and interviewer administration were only moderate (44%-78%) and there was a bias towards better retest (interviewer administration) scores on the physical fitness Chart.

Discussion

We used convenience samples of patients of a family medicine clinic because they were easily accessible and they represented the target population of the COOP/WONCA Charts. Our samples included males and females from different age groups and educational backgrounds, we believe that our results could be generalized to other Chinese adult patients in primary care.

The differences in the mean age, educational level and sex ratio among the three samples were expected, females and older people were less likely to be included in the two-method sample because more of them were illiterate. Any bias from the age and educational differences should have favoured the two-method sample who were younger and better educated, but this was not the case. Therefore, it was unlikely that these demographic differences had affected the reliability of the COOP/WONCA Charts.

We initially fixed the test-retest time-interval at 1 hour to be consistent with Nelson el al's study [1], but many subjects were unwilling to wait for an hour. We then allowed a flexible time interval between test and retest but the two must be separated by the doctor consultation. The relatively short time interval between test and retest could have inflated the reliability of the COOP/WONCA Charts but the interviewers did not find patients remembering their answers. This was supported by the fact that the concordance rates of the two-method sample were the lowest for most of the Charts although the mean test-retest time interval was the shortest.

Random replicative errors caused changes of no more than one in the COOP/WONCA Chart scores. A difference in the scores of one or more was likely to be a real difference if the COOP/WONCA Charts were administered by the same

interviewer. We found that the reliability coefficients of some of the COOP/WONCA Charts decreased with a change in the interviewer or administration method. When there was a change in the interviewer, a difference in the score of one could be the result of measurement variation although score changes of two or more were likely to be real. This implies that the health status of a patient could be monitored more reliably if there were personal continuity of care. On the other hand, one has to be aware of the tendency for patients to give more positive responses to some questions on repeated assessments by the same interviewer.

There was no significant bias in the retest scores when the Charts were administered by two different interviewers. This means that measurement variation would not cause any net change in the mean COOP/WONCA scores of a group of people. The Charts would be more reliable in detecting group differences than changes in an individual patient.

Our reliability coefficients were in general lower than those found in the US by Nelson et al[1]. Their reliability coefficients of the charts varied from 0.73 to 0.98 for the same interviewer and they were 0.50-0.98 for two interviewers. The reliability of the instrument might have been affected by the cultural tendency of the less educated Chinese to give socially approved answers as shown in an earlier survey with the Minnesota Multiphasic Personality Inventory 2 (MMPI-2) L scale [11,12]. We cannot assume that a health measure that has been shown to be reliable in one culture will be so in another. The reliability of an instrument must be confirmed on the target population before it is applied cross-culturally.

We found that the physical fitness and daily activities scores could differ by up to three when they were obtained by two different methods. Our interviewers

noticed that some subjects misinterpreted the physical fitness and daily activities charts as an assessment on what they actually did rather than what they could do. The meaning could be clarified when the Charts were administered by an interviewer but not when they were self-completed. This might be the reason why the scores obtained by self-completion were worse than those obtained by interviewer administration.

It is disturbing to find that self-completion and interviewing could give markedly different results. This is particularly relevant to family practice in that we often use the two methods together to collect patient information in clinical practice and research. Evidence on the technical equivalence of these two methods is few and conflicting. Some studies showed that there was little difference but others found that interviewer administration was more reliable[13]. Our study also showed that a change in the method of administration affected some results but not the others, probably because some questions were more prone to misinterpretation. Self-completion is more liable to give missing, inconsistent or inaccurate data, but an interviewer may be a barrier to honest responses. One method may be more suitable than the other for certain types of information. The effect of the method of data collection on the quality of information deserves more attention and research.

Conclusions

The COOP/WONCA Charts were reliable in detecting true differences between Chinese subjects when they were administered by the same interviewer. The reliability decreased but it was still within acceptable standard when the Charts were administered by different interviewers. The reliability of three Charts was quite low when they were administered by both self-completion and an interviewer. Misinterpretation of the questions could be a problem in self-completion of the Charts. Interviewer administration is the method of choice when the COOP/WONCA Charts are applied to the Chinese until we have more data confirming the reliability of self-completion.

We recommend the use of a single interviewer in the administration of the COOP/WONCA Charts to the Chinese if it is possible. When more than one interviewer are used, one must be aware of the inter-observer errors and differences in scores of less than two need be interpreted with caution. Self-completion and interviewer administration could give very different results for the same individual, the two methods should not be used together in the same survey and it may not be appropriate to compare data collected by different methods.

We found that a change in the method of administration caused significant changes in the COOP/WONCA scores despite the simplicity of the instrument. The method of administration may have even a greater effect on the results of longer and more complex health surveys. The reliability of any instrument and method of administration need to be confirmed on the target population before they are applied to clinical practice or research, otherwise, the results could be misleading. This is particularly important when cross-cultural adaptation is necessary.

References

1. Nelson E, Wasson J, Kirk J, et al. Assessment of function in routine clinical practice: description of the COOP chart method and preliminary findings. J Chron Dis 1987; 40 (Suppl 1): 55S-63S.
2. Scholten JHG, Van Weel C. Functional Status Assessment In Family Practice. Lelystad: Meditekst;1992.
3. Van Weel C, Konig-Zahn C, Touw-Otten FWMM, Van Duijn NP, Meyboom-de Jong B. Measuring Functional Health Status With The COOP/WONCA Charts. Groningen: Northern Centre of Health Care Research (NCH); 1995. NCH series no 7.
4. Lam CLK, Van Weel C, Lauder IJ. Can the Dartmouth COOP/WONCA Charts be used to assess the functional status of Chinese patients? Fam Pract 1994; 11: 85-94.
5. Census and Statistics Department, Hong Kong. Hong Kong Social and Economic Trends 1982-1992. Hong Kong : Government Printer; 1993.
6. Asian Development Bank. Key Indicators of Developing Asian and Pacific Countries. Manila: The Bank; 1993: Vol. 24.
7. Kerlinger FN. Foundations Of Behavioral Research. 3rd edition. Orlando: Holt, Rinehart & Winston; 1986: Chapter 26.
8. Helmstadter GC. Principles Of Psychological Measurements. New York: Appleton Century Crofts; 1964: Chapter 3.
9. Mongomery DC. The Design And Analysis Of Experiments. 3rd edition. New York: Wiley; 1991.
10. Nunnally JC. Psychometric Theory. 3rd edition. New York: McGraw Hill 1994.

11. Butcher JN. Introduction to the MMPI-2. In: Butcher JN (ed). MMPI-2 in Psychological Treatment. New York: Oxford University Press, 1990: 5-20.

12. Lam CLK, Chan MS, Poon V. Health survey tools- what work and what don't? (abstract). In: Irish College of General Practitioners, People and Their Family Doctors- Partners in Care, Book of Abstracts of the 15th WONCA World Conference; 1998; Dublin, Ireland. Oxford: Alden Press, 1998: 247.

13. Cella DF, Lloyd SR, Wright BD. Cross-cultural instrument equating: current research and future directions. In: Spilker B (ed), Quality of Life and Pharmacoeconomics in Clinical Trials, 2nd edition. Philadelphia, USA: Lippincott-Raven, 1996: 707-715.

Acknowledgment

This study was funded by a research grant from the Committee on Research and Conference Grants, the University of Hong Kong. We would like to thank Ms. Cyrina Chan and Ada Au for helping us with the data collection.

Table 1: Subject Characteristics

	Two-interviewer (n=84)	Two-Method (n=195)	One-interviewer (n=208)
Females#	62%	70%	77%
Mean Age*	63 years	42 years	58 years
No schooling#	54%	0.5%	36%
School > 6 years#	14%	56%	33%
Test-retest Interval*	57 minutes	32 minutes	40 minutes

*** The difference between the means is statistically significant ($p<0.01$) by t test.**
 # The difference between the proportions is statistically significant ($p<0.01$) by Chi square test

Table 2: Variance and 95% Confidence Interval[a] of Score Changes of the COOP/WONCA Charts

	Variance (95% Confidence Interval of Score Changes)			
COOP/WONCA Charts	**Inter-subject Variance (V) (n=208)**	**Random Variance (Vr) (n=208)**	**Inter-observer Variance (Vo) (n=44)**	**Inter-method Variance (Vm) (n=195)**
Physical Fitness	1.30 (±2.280)	0.12 (±0.693)	0.16 (±0.800)	0.8 (±1.789)
Feelings	0.60 (±1.549)	0.14 (±0.748)	NS	NS
Daily Activities	0.73 (±1.709)	0.27 (±1.039)	0.15 (±0.775)	0.43 (±1.311)
Social Activities	0.56 (±1.497)	0.08 (±0.566)	0.13 (±0.721)	NS
Health Change	0.43 (±1.311)	0.20 (±0.894)	NS	NS
Overall Health	0.37 (±1.217)	0.13 (±0.721)	0.13 (±0.721)	NS

Statistical Notes

(i). *a= 95% confidence interval = ± 2 standard deviation = ± 2 x √Variance*

(ii). *V is the variance component purely due to differences between subjects after exclusion of intra-observer random replicative variance (Vr). Vo is the variance component purely due to differences between observers, after exclusion of intra-observer variance. Vm is the variance component purely due to difference between administration methods, after exclusion of Vr and Vo.*

(iii). *The variance components presented for V, Vr, Vo and Vm are significantly greater than zero by the variance ratio F test at the 5% significance level. NS denotes non-significant variance components.*

Table 3: Reliability Coefficients of the COOP/WONCA Charts by the number of observers/methods

COOP/WONCA Charts	Same Observer V/(V+Vr)	Two Observers V/(V+Vr+Vo)	Two Methods V/(V+Vr+Vo+Vm)
Physical Fitness	0.915	0.823	0.546
Feelings	0.811	0.811	0.811
Daily Activities	0.730	0.635	0.462
Social Activities	0.875	0.727	0.727
Health Change	0.683	0.683	0.683
Overall Health	0.740	0.587	0.587

Table 4: Paired Differences in the Test-retest Scores of the COOP/WONCA Charts

Number (%) of Subjects

	- 2+	- 1	0	+ 1	+ 2+
Physical Fitness					
same interviewer	2 (1.0)	13 (6.3)	181 (87)	11 (5.3)	1 (0.5)
two interviewers	5 (11.4)	6 (13.6)	28 (63.6)	4 (9.1)	1 (2.3)
two methods*	55 (28.2)	43 (22.1)	86 (44.1)	7 (3.6)	4 (2.0)
Feelings					
same interviewer*	6 (2.9)	21 (10.1)	174 (83.7)	7 (3.4)	0
two interviewers	1 (2.3)	10 (22.7)	26 (59.1)	6 (13.6)	1 (2.3)
two methods	10 (5.2)	31 (16.1)	121 (63.0)	24 (12.5)	6 (3.1)
Daily Activities					
same interviewer*	12 (5.7)	21 (10.1)	164 (78.8)	9 (4.3)	2 (1.0)
two interviewers	0	3 (6.8)	35 (79.5)	4 (9.1)	2 (5.0)
two methods	11 (5.7)	38 (19.6)	109 (56.2)	30 (15.5)	6 (3.1)
Social Activities					
same interviewer	4 (1.9)	6 (2.9)	190 (91.3)	7 (3.4)	1 (0.5)
two interviewers	1 (2.3)	1 (2.3)	38 (86.4)	3 (6.8)	1 (2.3)
two methods	8 (4.2)	25 (13.1)	131 (68.6)	21 (11.0)	6 (3.1)
Change Health					
same interviewer	5 (2.4)	24 (11.5)	158 (76.0)	17 (8.2)	4 (1.9)
two interviewers	2 (4.5)	4 (9.1)	34 (77.3)	3 (6.8)	1 (2.3)
two methods	2 (1.0)	15 (7.7)	152 (78.4)	17 (8.7)	8 (4.1)
Overall Health					
same interviewer	1 (0.5)	22 (10.6)	166 (79.8)	17 (8.2)	2 (1.0)
two interviewers	1 (2.3)	11 (25.0)	26 (59.1)	6 (13.6)	0
two methods	3 (1.5)	36 (18.5)	134 (68.7)	18 (9.2)	4 (2.0)

* $p<0.05$ by the Wilcoxon matched-pairs signed-ranks test.