Cross-sectional Comparison of Live and Interactive Voice Recognition Administration of the SF-12 Health Status Survey

Richard W. Millard, PhD, MBA; and Joseph R. Carver, MD

Abstract

Objective: **To compare interactive voice recognition (IVR) and live telephone methods for administering the SF-12 health status survey (SF-12).**

Study Design: **Patients with low back pain received either IVR or live interviews in a cross- sectional design with partial randomization. The interviews consisted of the SF-12 and some additional questions specific to low back pain.**

Patients and Methods: **Complete findings were obtainable from 229 patients. Summary scales were compared by using multivariate analysis of variance with mean comparisons for continuously scored items. Response frequencies for categorically scored items were compared by using the chi-square test.**

Results: **The 2 methods produced similar results on the Physical Component Summary scale but not the Mental Component Summary scale. Compared with patients who had a live telephone interview, the patients using IVR acknowledged significantly greater overall mental interference, greater general emotional concerns, and poorer mood and overall health.**

Conclusions: **Because IVR eliminates the demand characteristics of responding to a personal interviewer, it may be a desirable way to evaluate sensitive topics. It also may reduce costs of data entry, labor, and measurement error.**

(Am J Managed Care 1999;5:153-159)

Health status surveys are gaining widespread visibility as tools to measure patient experiences. In contrast to other kinds of clinical measurements, these instruments are designed for use with large populations, often as part of quality improvement initiatives or for documentation of clinical outcomes. Because of both increasing use of these instruments and a focus on cost-effective use of healthcare resources, economic collection of patient-generated health status surveys is a logical concern. The original prototype, the SF-36, is a "long" survey and was a product of the Medical Outcomes Study,[1] in which the development of health status measures was an explicit goal.[2] Its 36 questions provide a wealth of detail but require considerable time and resources for completion. Fortunately, briefer questionnaires that make fewer demands on the respondent and surveyor are available to gather information. The SF-12 is a validated short-form ("SF") instrument composed of 12 items contained in the SF-36. Like the SF-36, the SF-12 asks about physical activities/function (eg, using stairs), pain interference, and emotional status/distress (eg, feeling calm, feeling downhearted). Results are expressed as Physical Component Summary (PCS) and Mental Component Summary (MCS) scores. In contrast to the SF-36, the entire questionnaire can be completed in about 3 or 4 minutes.

The SF-12 has been subjected to a high level of psychometric scrutiny. Based on a general US population sample (n = 2333), R^2 values of >0.90 were reported when the SF-12 was used to predict SF-36 results.[3] Comparisons between many and diverse patient samples support the equivalence of the SF-12 to its longer parent.[4] In one study, the PCS and MCS scores from the two versions were virtually identical in patients with conditions as diverse as congestive heart failure, sleep apnea, or inguinal hernia.[5]

From Patient Infosystems and University of Rochester School of Medicine, Rochester, NY (R.W.M.); and Aetna US Healthcare, Blue Bell, PA (J.R.C.).

This research was supported by Patient Infosystems, Inc., Rochester, NY.

Address correspondence to: Richard W. Millard, PhD, MBA, Patient Infosystems, 46 Prince St, Rochester, NY 14607. E-mail: rmillard@ptisys.com.

Reducing the number of questions from 36 to 12 leads to a corresponding reduction in the time required for administration. Once the issue of form length (long vs short) was resolved, the next logical approach to evaluating efficiency and cost-effectiveness was to alter the route of administration by using the telephone and computer technology. Examples include interactive voice recognition technology (IVR), optical scanning, software to simplify scoring procedures, and reports that are generated via fax. We evaluated IVR technology compared with traditional live telephone interviews as a method to administer the SF-12.

IVR is an automated intervention that uses specialized telephone hardware and the manipulation of digitized voice. The voice that is heard during this kind of encounter may be either prerecorded or computer synthesized. The respondent's spoken answer is automatically recognized by this system. Early versions of IVR made use of touch tone keypad entries in lieu of a verbal reply. The use of IVR technology has been well documented. For example, 1812 respondents used this technique in a study of a call-in screening service for depression.[6] Preliminary evidence suggests that subjects may actually provide more honest responses, particularly with respect to substance abuse, when using IVR systems compared with live telephone interviews.[7-9] IVR administration presumably reduces bias associated with volunteering information that is socially undesirable instead of purely factual.

IVR is increasingly available to conduct health status surveys, although the advantages of using this data collection method are not fully understood. It is important to determine whether the IVR approach results in the same kinds of measurements that are obtained with other methods. Using a sample of patients with chronic low back pain, we sought to compare the SF-12 results obtained by IVR with those obtained by live telephone interviews. The study design attempted to randomize assignment to the two conditions of survey administration. First, it was hypothesized that the 2 methods would result in similar PCS scores. At the same time, we anticipated a trend toward reports of greater mental interference (MCS score) among patients who responded through use of the IVR system. If a pattern of differences in reported mental interference was found, then it was hypothesized that these differences would be most apparent for those SF-12 questions with more sensitive psychosocial content.

··· METHODS ···

The patients who participated in this study were undergoing health status assessment for low back pain, conducted as part of the Healthy Outlook Program® for Aetna US Healthcare (Blue Bell, PA). Patients completed the SF-12 by use of either IVR or a live telephone interview. The projected minimum sample size was determined via power analysis (effect size = .4, β = .8, α = .05), with at least 50 patients sought for each of the assessment conditions.[10] Live and IVR interviews were conducted at the Health Information Call Center of Patient Infosystems, which was acting as an agent for Aetna US Healthcare. Demographic and telephone data on patients were transferred via electronic data feed from Aetna US Healthcare to Patient Infosystems. The records for each patient were then transferred into a queuing software program so that the patient's telephone number could be assigned to 1 of 14 patient service representatives (PSRs) to complete the telephone calls. Calls were automatically dialed.

A queuing program routed each call to the next available PSR and functioned to randomize assignment to each of the PSRs. Each successive patient in the queue had an equal probability of being assigned to each of the PSRs, who were instructed to alternate IVR and live administrations. Perfect randomization was not obtained because some patients who were assigned to the IVR condition exhibited cognitive deficits or speech or hearing difficulties that made it necessary to complete their calls via live interview.

Telephone interviews were conducted so that a PSR would speak personally with each patient at the beginning of the call. This interval, usually lasting about 1 minute, was used to verify the patient's identity and to confirm his or her willingness to complete the survey. During this time, the PSR also could evaluate whether the patient had speech or hearing impairment or cognitive difficulties (in terms of poor comprehension or long response latencies); these were specific exclusion criteria for the IVR format. PSRs were instructed to alternate between live and IVR administration for each call that they received during the study period, which consisted of two 6-hour blocks of time approximately 1 month apart. In this way, each PSR handled both live and IVR interviews. For live interviews, they made use of a script for administration of the SF-12.[4] The procedure for patients who completed the IVR interview also began with live con-

tact with a PSR, who explained use of the speech recognition system. The patient then was transferred to the IVR system to complete the SF-12, with the same script as for live administration. Patients always had the option at any point during the IVR interview to return to the PSR. If a patient did not answer a question in the IVR format, it would be repeated twice before being recorded as a missing response. If 2 questions were not answered, the call would automatically revert to the PSR.

PSRs completed 50 hours of training to demonstrate their ability to conduct interviews in the health information call center setting. This training included role playing and ongoing covert monitoring of their performance during interviews. They were fully instructed to adhere to the use of standard scripts.

In addition to the questions on the SF-12, the patient interview included a series of questions about interference due to back pain. These 2 components were not timed separately. The results of these additional questions were not compared in this study because they did not constitute a standard scale.

The planned analyses consisted of multivariate analysis of variance (MANOVA) followed by comparisons of mean PCS and MCS SF-12 scores between the 2 conditions (live vs IVR interviews). Further comparisons were planned between each of the 12 items on the SF-12, conditional on demonstrating differences in summary scale results. For the 8 continuously scaled responses, MANOVA was used, followed by comparison of means. Chi-square frequency comparisons were used to compare responses on the 4 categorical items.

··· RESULTS ···

During the study period 288 patients completed SF-12 interviews. More patients completed interviews during the study period than was anticipated. As a result, the sample size exceeded what had been identified as necessary during power analysis. The records for 46 patients were not appropriate for analysis. The most common reason was because the patient elected to return to the PSR during the call (27 patients), which resulted in combined IVR and live SF-12 administration. The IVR and live administration groups were equivalent with respect to patient gender (IVR: 53 women and 54 men; live: 66 women and 69 men).

Patients who used IVR were younger (mean age = 56.96 years, SD = 14.87 years) than those who used live administration (mean age = 62.61 years, SD = 14.79 years) (t = 2.87, P < .05). This result, which partially violated randomization, was presumably attributable to older patients more frequently having health problems (eg, hearing loss) that required live administration. Because of this difference, age was used as a covariate in further analyses. This was done so that any apparent differences in response to IVR vs live interviews were not attributable to age-related effects. The mean age of the 27 patients who returned to a live interviewer during the IVR portion was 60.66 years (SD = 11.18 years). Live interviews took slightly longer to complete and were more variable in duration than IVR interviews (IVR: mean time = 10.6 minutes, SD = 1.94 minutes; live: mean time = 11.36 minutes, SD = 4.14 minutes). These times include both the SF-12 and disease-specific questions about low back pain. The difference in call duration was not significant (Student t test).

Some patients did not provide answers for every question of the SF-12. Patients might not answer a specific SF-12 question for various reasons, for example, because it did not seem relevant or because it was difficult to understand. IVR made use of a "beep" prompt, which signified when the patient was expected to answer questions. Failing to wait for this prompt would be recorded as no answer if it occurred once. If it occurred twice, the call would be transferred back to the PSR. To make the comparisons more meaningful, patients with missing responses were removed from further analyses. This reduced the sample size to 229 patients: 98 patients who completed the IVR interview and 131 who completed the live telephone interview.

In summary, the reasons that fewer patients completed the SF-12 using the IVR format were because a cognitive, hearing, or speech impairment was detected by the PSR on initiating the call, because the patient elected to return to the PSR during the call, or because of missing data. Apart from age and gender, no further demographic information was available to indicate which variables distinguished the sample of patients who completed the entire interview in the IVR format from those who completed the interview with assistance from a PSR. By using age as a covariate in the analyses, it was presumably possible to control for some of the variance that would be attributable to age-related factors affecting participation in the IVR system.

Unadjusted mean scores and standard deviations for live and IVR ad-ministrations are presented in

Table 1. SF-12 Summary and Item Scores Before Adjustment for Age

Summary Scale or Item	Live (n=131)	IVR (n=98)
Mean Score (SD), Summary Scales		
PCS*	40.26 (10.63)	38.63 (11.09)
MCS*	48.80 (12.01)	43.81 (8.76)
Mean Score (SD), Continuous Items		
Overall health	2.71 (1.41)	2.95 (1.11)
Moderate activities*	2.66 (1.19)	2.66 (1.04)
Using stairs*	2.25 (1.04)	2.37 (0.88)
Pain interference	2.92 (1.40)	3.01 (1.40)
Social interference	2.21 (1.50)	2.47 (1.44)
Calm and peaceful	2.87 (1.50)	3.30 (1.53)
Lot of energy	3.45 (1.53)	3.61 (1.40)
Downhearted and blue*	4.16 (1.71)	3.22 (1.76)
Mean Score (SD), Categorical Items		
Accomplished less, physical	0.60 (0.49)	0.65 (0.48)
Limited work/activities, physical	0.53 (0.50)	0.61 (0.49)
Accomplished less, emotional	0.21 (0.41)	0.38 (0.49)
Limited work/activities, emotional	0.21 (0.41)	0.26 (0.44)

PCS =Physical Component Scale; MCS = Mental Component Scale; IVR = interactive voice recognition.
*Higher scores reflect less interference.

Table 2. PCS and MCS Results Correlated to Patient-Reported Pain, Symptom Understanding, and Interference*

	Live		IVR	
Variable	**PCS**	**MCS**	**PCS**	**MCS**
Understand what makes symptoms worse	–.00	.06	–.19	.09
Pain constancy	–.50§	–.30‡	–.51§	–.29‡
Days of interference per month	–.61§	–.33‡	–.52§	–.20+
Days of missed work per month	–.32‡	–.21+	–.41§	–.25+

PCS = Physical Component Scale; MCS = Mental Component Scale; IVR = interactive voice recognition.
*Higher scores reflect less interference.
+$P < 0.05$
‡$P < 0.01$
§$P < 0.001$

Table 1. The greatest disparity occurred for the question about being "downhearted or blue." Internal consistency, as measured by Cronbach's alpha, was high overall; $r = 0.85$ for the total sample (n = 229). This value was similar for the sample of patients using IVR ($r = 0.87$, n = 98) and those completing live interviews ($r = 0.83$, n = 131). Because the PCS and MCS are actually weighted composite values, Cronbach's alpha was calculated for the 12 items instead of the summary scales. The standard errors of measurement were calculated for PCS and MCS across conditions. These values were similar for PCS scores (live = 46.84, IVR = 44.35). There was more variability in MCS scores for the patients interviewed live (live = 59.80, IVR = 27.67).

The PCS and MCS scores obtained with the 2 methods of administration were similarly associated to other clinical variables. This information is summarized in Table 2. With either type of administration, physical interference and pain constancy were more highly associated with PCS scores than with MCS scores. The correlation to patient-reported understanding of symptoms was very low for both administration methods. Age showed a modest but significant correlation to MCS results in the IVR sample ($r = 0.21$, $P < .05$) and lower, nonsignificant correlations to MCS results in the live sample and to PCS results in both samples.

MANOVA results showed a significant overall effect when comparing PCS and MCS scores for the two forms of administration (Wilks lambda = 0.96, F = 4.77, $P < .01$). The MANOVA used age as a covariate, resulting in 2 degrees of freedom for the numerator. In other words, the procedure was like a mixed-model analysis of variance evaluating a method by subscale interaction, except that age served as a covariate. There was a trend toward greater report-

ed interference with the IVR format. Mean comparisons showed that the difference between PCS scores was not significant. However, patients who had the IVR interview did report significantly greater mental interference as measured by the MCS score ($F = 8.72$, $P < .01$).

The second MANOVA concerned differences between the two samples in terms of patient responses to the 8 continuously scored SF-12 items. (Response choices were continuously distributed from 0 or 1 to 5 or 6.) This overall model also was significant (Wilks lambda = 0.90, $F = 2.97$, $P < .01$). The comparisons between live and IVR responses are displayed in Table 3. There were significant differences in responses to 2 items. IVR respondents indicated that their "overall health" was poorer than patients who completed the SF-12 as a live interview. In addition, the IVR respondents reported being more frequently "downhearted or blue."

A separate phase of analyses was conducted to compare responses to the remaining 4 SF-12 questions, which require categorical responses. These comparisons were done by using chi-square analysis. As categorical data in a nonparametric analysis, these data were not age adjusted. The results are summarized in Table 4. A significant difference was present for one item, which asked whether a patient "accomplished less . . . as a result of any emotional problems, such as feeling depressed or anxious." Patients who answered through IVR were significantly more likely to answer this question affirmatively.

Table 3. Comparison of Least Square Mean Scores for SF-12 Summary Scales and Continuously Scored Responses

Summary Scale or Item	Live (n = 131)	IVR (n = 98)	F Value *df* (2,225)
Mean Score, Summary Scale			
PCS*	40.22	38.68	1.09
MCS*	48.50	44.22	8.72[+]
Mean Score, Continuous Items			
Overall health	2.69	2.98	3.81[‡]
Moderate activities	2.65	2.67	0.02
Using stairs	2.22	2.41	1.99
Pain interference	2.95	2.98	0.03
Social interference	2.26	2.43	0.72
Calm and peaceful	2.91	3.26	2.82[+]
Lot of energy	3.47	3.59	0.33
Downhearted and blue*	4.14	3.25	14.58[§]

PCS = Physical Component Scale; MCS = Mental Component Scale; IVR = interactive voice recognition.
*Higher scores reflect less interference.
[+]$P < .01$.
[‡]$P < .05$.
[§]$P < .001$.

Table 4. Comparison of Response Frequencies for Categorically

Item	Yes (%)		χ^2
	Live	IVR	
Accomplished less, physical	61.48	68.22	1.19
Limited work/activities, physical	54.48	63.55	2.02
Accomplished less, emotional	20.74	36.45	7.36*
Limited work/activities, emotional	20.00	27.10	1.69

Scored SF-12 Items
IVR = interactive voice recognition.
*$P < .01$.

··· DISCUSSION ···

The first hypothesis, that IVR and live interviews would yield similar SF-12 PCS scores, was substantiated in this cross-sectional sample. It would appear that the absence of a human interviewer does not introduce a significant level of additional measurement error or bias for questions about physical activities. However, the similarities in results are probably less meaningful than the differences that occurred. On the MCS score, there was significant evidence of greater reported mental interference among patients who used the IVR method.

Although there was an overall trend toward reporting greater interference when using the IVR format, patient responses were significantly different when IVR was used to determine their mood state. This pattern was most evident with the item that asked about sadness or depressed mood. There was also a significant difference in acknowledging emotional concerns (without age adjustment), but not specifically with reference to work or regular activities. These results are consistent with the second hypothesis, that differences would be most apparent for items with sensitive emotional content. In addition to greater reported interference in these areas, IVR respondents also described more negative judgments about their current health and daily activities. Such differences did not emerge in comparing questions about energy level, calmness, or physical health. It might be argued that the demand characteristics of responding to a personal interviewer may introduce a source of distortion not present during IVR administrations. This issue deserves more complete consideration in subsequent research, perhaps by systematically controlling for interviewer characteristics.

IVR is still a very new way to obtain clinical measurements, so few studies have investigated this topic. Kobak et al. reported that IVR has greater sensitivity in detecting alcohol problems, but not other psychiatric conditions.[8] A study of adolescent sexual behavior and drug abuse also found higher prevalence rates when IVR was used.[7] The present results are consistent with this pattern of improved sensitivity in detecting emotional concerns. In contrast, well-designed comparisons of more conventional administration methods have generally indicated equivalence, for example, between telephone and in-person interviews (patients with bipolar disorder)[11] and between questionnaires and in-person interviews (patients with AIDS).[12]

Because these data were cross-sectional rather than longitudinal, repeat crossover administrations are being planned. A second phase will involve having patients complete the SF-12 by using an alternate form, either IVR or live, at a second administration. In that way individual subjects' responses can be compared over time, ideally with short (eg, 2-5 days) intervals between administrations.

A further constraint was imposed by the lack of descriptive demographic information about patients who did not complete the interview in the IVR format. Clearly, the IVR approach is not workable for all patients. Even with initial screening, some patients do not complete interviews in the IVR format. The 27 patients who switched from IVR to live administration after beginning their call (excluded from analyses) might have done so to ask for clarification about the meaning of a question or because they had incorrectly registered responses (reasons for discontinuation were not classified for live interviews). They were slightly older than patients who completed IVR but younger than those who completed the live interview. The patients who discontinued IVR were more frequently women than men (63% female), whereas there were equal proportions of women and men in the 2 samples that completed their interviews.

Better knowledge of individual differences would assist in identification of factors that might affect participation in the automated system. Other studies are currently in progress with our group to establish a profile of which variables (eg, education, diagnosis, voice quality) are predictive of difficulty in using the IVR system. This information would be used to supply PSRs with more detailed criteria about when to conduct an interview entirely in a live format. Preliminary evidence with a sample of patients with congestive heart failure has suggested that the most frequent reason that patients discontinue IVR is because they fail to wait long enough to register their answer once the question is stated, not because of technological failings in recording accuracy. Although IVR program applications use a beep to signal when it is time to make a response, not all patients wait for this sound before answering. The complexity of the questions being asked is also likely to affect participation rates, although this topic has not been systematically explored.

It also would be appropriate to conduct further comparisons among other diagnostic groups to evaluate whether similar response patterns are present for other health conditions. Chronic low back pain is a health problem with significant affective fea-

tures. The prevalence of major depression is thought to be up to 4 times greater in these patients than in general population, exclusive of mild mood disturbances that may alter symptom perception.[13] Patients with a high level of somatic preoccupation may be reluctant to fully acknowledge mood concerns for fear that this acknowledgement will result in ostracism or inattention to physical processes. This may help to explain why a more anonymous reporting format resulted in greater reports of mental interference in this clinical sample. In comparison to published SF-36 norms for the general US population with back pain or sciatica, this sample reported slightly greater physical interference but comparable levels of mental interference (statistical comparisons, however, were not performed).[14]

··· CONCLUSION ···

If IVR facilitates the collection of less biased clinical data, that would be a compelling reason to choose it. Advantages include real-time data compilation, less need for human labor, and less possibility of human error. Ideally, this tool would make it possible to reach patients proactively, rather than waiting for emergencies to trigger clinical encounters. Shorter health status tools such as the SF-12 can be easily implemented in an IVR format. With the advent of more sophisticated evaluation algorithms that use branching, adaptive logic, there should be further opportunities to use and evaluate this kind of technology as a way to efficiently gather accurate patient data.

Acknowledgements

We thank Kathleen Holt, PhD, for valuable assistance conducting statistical analyses and Hodge Griffone and Saby Kulkarni for supervising survey administrations.

··· REFERENCES ···

1. Ware JE, Sherbourne CD. The MOS 36-item short form survey (SF-36), I: Conceptual framework and item selection. *Med Care* 1992;30:473-481.
2. Tarlov AR, Ware JE, Greenfield S, et al. The medical outcomes study. An application of methods for monitoring the results of medical care. *JAMA* 1989;262:925-930.
3. Ware JE, Kosinski M, Keller SD. A 12-item short-form health survey. Construction of scales and preliminary tests of reliability and validity. *Med Care* 1996;34:220-233.
4. Ware JE, Kosinski M, Keller SD. SF-12: *How to Score the SF-12 Physical and Mental Health Summary Scales.* Boston, MA: The Health Institute; 1995:43-58.
5. Jenkinson C, Layte R, Jenkinson D, et al. A shorter form health survey: Can the SF-12 replicate results from the SF-36 in longitudinal studies? *J Public Health Med* 1997;19:179-186.
6. Baer L, Jacobs DG, Cukor P. Automated telephone screening survey for depression. *JAMA* 1995;278:1943-1944.
7. Turner CF, Ku L, Rogers SM, et al. Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science* 1998;280:867-873.
8. Kobak KA, Taylor LH, Dottl SL, et al. A computer-administered telephone interview to identify mental disorders. *JAMA* 1997;278:905-910.
9. Alemagno SA, Cochran D, Feucht TE, et al. Automated monitoring of outcomes: Application to treatment of drug abuse. *Am J Public Health* 1996;86:1626-1628.
10. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. New York, NY: Lawrence Erlbaum; 1988.
11. Revicki DA, Tohen M, Gyulai L, et al. Telephone versus in-person clinical and health status assessment interviews in patients with bipolar disorder. *Harv Rev Psychiatry* 1997;5:75-81.
12. Wu AW, Jacobson DL, Berzon RA, et al. The effect of mode of administration on medical outcomes study health ratings and EuroQOL scores in AIDS. *Qual Life Res* 1997;6:3-10.
13. Sullivan MJ, Reesor K, Mikail S, Fisher R The treatment of depression in chronic low back pain: Review and recommendations. *Pain* 1996;50:5-13.
14. Ware JE, Kosinski M, Keller SD. *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Boston, MA: The Health Institute; 1996.