Final Report

HOS/VA Comparison Project

Part 2: Tests of Reliability and Validity at the Scale Level for the Medicare HOS MOS -SF-36 and the VA Veterans SF-36

Lewis E. Kazis, Austin F. Lee, Avron Spiro III, Donald R. Miller, William Rogers,
Xinhua S. Ren, Milanda Zhang.

Health Outcomes Technologies Program
Health Services Department
Boston University School of Public Health

Contents

The Centers for Medicare & Medicaid Services' Office of Research, Development, and Information (ORDI) strives to make information available to all. Nevertheless, portions of our files including charts, tables, and graphics may be difficult to read using assistive technology.

Persons with disabilities experiencing problems accessing portions of any file should contact ORDI through e-mail at ORDI_508_Compliance@cms.hhs.gov

0. Abstract

Objective: This report provides psychometric evidence for the comparability of the V/SF-36 and MOS SF-36 for potential use in future studies comparing outcomes in different healthcare systems. The two SF-36 questionnaires are distinctly different in terms of the role physical and role emotional scales and the component summaries. The objective is to examine the reliability and discriminant validity of the V/SF-36 and MOS SF-36 versions on a scale by scale and summary level, for the physical and mental summaries.

Methodology: The sample consists of 4,528 persons who responded to both the 1999 Large Health survey of Veteran Enrollees (VA Survey) and Cohort 2 (1999) of the Health Outcomes Survey (HOS). We also merged the clinical diagnoses using ICD-9CM codes from the VA data sets. Of these 4,528 cases, SF-36 scores could be computed for 2,737 (60%). Analysis involves assessment of reliability using Cronbach's Alpha, multitrait scaling, and factor analysis using principal iterations and varimax rotation, and discriminant validity testing using a well-validated index of comorbidities.

Results: The V/SF-36 yielded consequential improvements over the MOS SF-36 in terms of Cronbach's Alpha reliability for the role physical and role emotional scales, (0.96 versus 0.91, and 0.95 versus 0.89, respectively). Improvements to the precision of the scales are particularly marked for the role physical and role emotional scales. The multitrait scaling analyses for these two role scales also demonstrated greater internal consistency at the item level for the V/SF-36 than the MOS SF-36. The factor analysis strongly suggested that the scales are comparable for the two versions. Both forms of the SF-36 demonstrated adequate discriminant validity using the number of medical and mental comorbidities derived from ICD-9 CM Codes from the VA. However the role physical and emotional scales from the V/SF-36 version demonstrated a lower floor and greater efficiencies than the MOS version, with the exception of the number of mental comorbidities for the role-emotional scale which were almost comparable (less than 5% difference in efficiency). For the physical (PCS) and mental (MCS) summaries, we note a marked effect on the floor of the scale for the V version which is quite a bit lower for MCS and slightly lower for PCS when compared to the MOS version. We also report substantial gains in the efficiency of the V version for the MCS summary for the number of medical and mental comorbidities.

Conclusions: The V/SF-36 scales and component summaries are at least as reliable and valid as the MOS version, and in fact are improved for the role scales and summaries. The results strongly suggest that the V/SF-36 is suitable for comparisons at the scale level with the MOS version. The gains in the precision in the V/SF-36 for the two role scales are quite important and provide evidence for the use of the V/SF-36 in future applications for assessing outcomes of health care systems.

1. Overview and Objectives

Part 1 of this report provided psychometric evidence at the item level for the comparability of the V/SF-36 and MOS SF-36. With this comparability established, we can conclude that the two versions can be used in future studies to compare outcomes between healthcare systems.

The objectives of part 2 of this report are to:

1. Examine the reliability and validity of the V/SF-36 and MOS SF-36 at the scale level using Cronbach's Alpha Statistics, multitrait scaling and factor analysis.

2. To examine the discriminant validity of the V/SF-36 and MOS SF-36 at the scale level using a measure of medical and mental comorbidities based on ICD-9 CM codes derived from VA administrative data.

The results of this report (on the reliability and validity of the V/SF-36 and MOS SF-36 at the scale level) can be used to determine whether and how to use these two versions for future system (Medicare versus VA) comparison studies.

2. Methods

2.1 Samples

This study uses data from the 1999 Large Health survey of Veteran Enrollees (VA Survey) (Perlin and Kazis et al. 2000) and Cohort 2 (1999) Health Outcomes Survey (NCQA 1998). The details of these two surveys are described in Part 1 of this report. Cohort 2 was chosen as it is most proximal in time to the VA survey. The VA survey was conducted from July 1999 – January 2000 and the HOS survey was fielded in March of 1999.

Briefly, there are 248,484 respondents in HOS cohort 2, and 887,775 respondents to the VA survey. After merging the two surveys (HOS and VA), there were 4,528 respondents who completed both the HOS and VA surveys. Of the 4,528 cases, 2,737 (60%) had sufficient data to compute SF-36 scores for the HOS and VA. Thus, for these 2,737 respondents, we computed both MOS SF-36 and V/SF-36 scale scores (for 8 scales) and component summaries (for physical and mental health).

Data for these respondents who were in both surveys then were merged with VA administrative data from the Outpatient (OPC) and Inpatient (PTF) files which include ICD-9 CM codes. These codes are fairly complete and provide diagnostic information for the three years prior to the VA survey (Perlin and Kazis et al. 2000).

2.2 SF-36

The MOS SF-36 is well documented and described elsewhere. The V/SF-36 version has been previously documented as reliable and valid in ambulatory VA patient populations, and has been adopted by the VHA as one of the performance measures of functional status (Kazis et al. 1999, 2000, 2003). It builds on the MOS SF-36 (Ware et al. 1992), and modifications include changes to the role items (role limitations due to physical and emotional problems). In particular, response choices that were originally dichotomous (yes/no) are now five-point ordinal choices ('no, none of the time' to 'yes all of the time').

Previous work has shown that these changes to the SF-36 increased the precision and discriminant validity of the role scales and physical and mental component summaries. Like the MOS version of the SF-36, the V/SF-36 measures eight concepts of health: physical functioning (PF), role limitations due to physical problems (RP), bodily pain (BP), general health perceptions (GH), energy/vitality (VT), social functioning (SF), role limitations due to emotional problems (RE), and mental health (MH). The V version of the SF-36 also includes two items dealing with patient perceived changes in physical and mental health over the past year, while the MOS version includes one overall item dealing with changes in health. The transition items are not included in this analysis. Conversion formulas have been developed and validated for scoring the V version role scales so that scores are comparable to the MOS version (Kazis et al. 2003, in press).

Items from each concept are summed and rescaled with a standard range from 0 to 100, where 100 denotes the best health. These eight concepts have also been summarized into two summary scores: *a physical component summary (PCS) and a mental component summary (MCS)* (Ware et al. 1993, 1994). The summary scales are based upon the finding that more than 90% of the reliable variance in the eight SF-36 scales are explained by the physical and mental dimensions of health. The reliable variance is higher in the V version than the MOS version (Kazis et al. 1999). As in the MOS version of the SF-36; the two component summary scales are each scored using weights derived from a national probability sample of the U.S. population. They are standardized to the U.S. population and norm-based so that the scores have a direct interpretation in relation to the distribution of scores in the U.S. population with a mean of 50 and a standard deviation of 10. Higher scores indicate better health. Each summary is expressed as a T score, which facilitates comparisons between the VA patients and the general US population. For the V/SF-36 the PCS and MCS scores are computed and the two summaries make an important contrast between the physical and psychological health status of veteran users of the VHA. The calculation of the scales and VA norms have been published and disseminated VA wide in each of five national survey reports from 1996 to 2000, representing close to 2 million administrations of the Veterans SF-36 (Kazis et al. 1997, 1998a, 1998b), (Perlin and Kazis 2000).

The Veterans SF-36 has been previously validated in the Veterans Health Study (VHS), with Cronbach’s Alphas ranging from 0.93 to 0.78 for physical and social functioning, respectively (Kazis et al. 1998, 1999). Published work from the VHS has demonstrated the discriminant validity of the scales and component summaries. The

Veterans SF-36 scores are strongly correlated with sociodemographics and morbidities of the veteran users of the VHA system of care (Kazis et al. 1998, 1999).

2.3 Psychometric Methods

Because the differences in the V/SF-36 and MOS SF-36 are in the role scales (role physical and role-emotional), we focus on these two scales and on the physical and mental summaries. However, for completeness, we also report on the results of the other 6 scales in the tables.

2.4 Cronbach's Alpha Statistics

Cronbach's alpha for a given scale is a function of the number of items and their average intercorrelation. This statistic is a measure of the precision of the measure. We generated Cronbach's Alpha statistics for each of the 8 scales of the V/SF-36 and MOS SF-36. We also report the reliability of the two component summaries PCS (physical summary) and MCS (mental summary). Because the component summaries are linear combinations of the eight scales, the reliability coefficient must take into account the reliability of each scale and the covariances among them using the internal consistency method (Ware et al. 1994, page 5:2). The measurement variance is based on a fundamental theorem about variances:

$$\text{Variance } (aX + bY) = a^2 \text{ Var}(X) + b^2 \text{ Var}(Y) + 2ab\text{Cov}(X,Y)$$

Since the scales are statistically independent (from a measurement viewpoint), the Covariance term drops away, and we can simply add the variances of the scales, multiplied by the square of their weights. The variance of the scale is (1-alpha) (ordinary scale SD)^2 and the weights are derived from the formulas for constructing PCS and MCS.

2.5 Multitrait Scaling

Multi-trait scaling uses convergent and discriminant validity to test the performance of items in their hypothesized scales. Item-scale correlations are the primary elements of multi-trait scaling (Hays et al. 1990). First, item internal consistency is assessed by determining if each item in a scale is substantially linearly related to the total score computed from other items in that same scale. Second, the item discriminant validity criterion is assessed by determining if each item has higher correlations with the scale it is hypothesized to belong to than with all other scales. These two tests gauge the consistency of items in their scale and their divergence from other items in different scales.

Item internal consistency is supported if an item correlates substantially ($r \geq 0.40$) with the scale it is hypothesized to represent. To correct for overlap, the hypothesized item is deleted from the scale with which it is correlated. Item discriminant validity depends upon the magnitude of the correlation between an item and its scale relative to the correlation of that item with other scales. If the hypothesized correlation is more than 2 standard errors higher than the other correlations a "scaling success" is counted, if it is more than 2 standard errors lower a "definite scaling error" is counted, and if it is within 2 standard errors of all correlations with other scales, it is considered a probable scaling error.

As already noted, to test for internal consistency, reliability coefficients (Cronbach's Alpha) will be computed for each of the scales, as well as the range of the correlations for both item internal consistency and item discriminant validity. Thus we will also include for the item discriminant validity testing the number of successes, the number of failures, and the number of probable failures for each of the 8 scales of both the V/SF-36 and the MOS SF-36.

2.6 Factor Analysis

Factor analysis is conducted for the eight scales for both the V/SF-36 and MOS SF-36, using principal iterations and varimax rotation. Eigenvalues are set to 1 prior to factor extraction. Both the variance explained by the rotated factor structure and communalities are reported for each. Comparisons are made between the V and MOS versions of the SF-36 based upon factor loadings, variance explained by the rotated factor structure, and communalities for the respective scales.

2.7 Discriminant Validity Testing of Scales by Clinical Group Comparisons:

Discriminant validity testing of the V/SF-36 and MOS SF-36 scales was conducted by comparing scale score means and standard deviations across groups of patients defined of different levels of clinical severity (as defined by the number of comorbidities). In this analysis, we assess the ability of the V/SF-36 and MOS SF-36 scales and summary scores (physical and mental component summaries) to discriminate among the groups stratified by a comorbidity index. This index is based upon ICD-9 CM codes obtained from the VA data base. The medical comorbidity index is a sum of medical conditions and can range from 0 to 30, while the mental comorbidity index can range from 0 to 6. Both are simple sums of conditions based on ICD-9 diagnoses obtained from VA administrative data over the 3 years prior to the VA survey. This comorbidity index, with its medical and mental indices, have been validated previously in the Veterans Health Study (Selim et al. 2003).

Analytic methods for assessing discriminant validity include general linear model procedures (ordinary least square regression) with the F statistics and associated p-values reported for the V/SF-36 and MOS SF-36 by scale and physical and mental summaries. The F statistic is compared for the V and MOS SF-36 versions. The F statistic is based upon the interaction term of the survey (V/SF-36 versus MOS SF-36) by the number of medical or mental comorbidities. We view this as a measure of the difference between the two measures ability to discriminate across the summative levels of comorbidity. This is a direct comparison of the trends of the two versions. A significant F statistic may be driven by the range of the differences of the scale scores.

Efficiency of the V/SF-36 and MOS SF-36 versions for role limitations due to physical problems (RP) and role limitations due to emotional problems are given by the ratio of the F statistics for each using one way analysis of variance. The ratio is computed relative to the MOS version (V version result on the numerator and MOS version in the denominator). We conduct a similar test of efficiency of the two versions using the physical and mental summaries (PCS and MCS). This approach for characterizing efficiency of the scales is well documented in the literature. We report on the differences in efficiency of RP, RE and PCS and MCS for the V and MOS versions of the SF-36.

2.8 Other Analytic Considerations (precision & ceiling/floor effects):

We also examined the range of mean scores across the levels of comorbidity, comparing scores between those with no comorbidities (minimal disease burden) and those with many comorbidities (great deal of disease burden), for the role scales and the physical and mental summary scores. Although we do not anticipate any differences between the other scales, we will examine them as well. Any differences that we do find may be the result of differences in the nature of the administrations for the VA and HOS surveys. We are particularly interested in identifying differences in floor and ceiling between the two versions, e.g., whether the V/SF-36 version has reduced floor effects compared to the MOS version for the role scales, as reported in previous work (Kazis et al. 2003).

2.9 Other Method Issues

The following analysis and report of results are based upon 4,528 subjects who responded to both the 1999 HOS (Cohort 2) and to the 1999 VA survey. Of these 4,528, SF-36 scores were complete for 60% (N= 2,737). We were very conservative in our approach for dealing with missing values and used the 50% rule (i.e., if more than 50% of the items for a given scale or concept were missing, then we coded the scale as missing). More in- depth analyses using advanced imputation methodologies are planned for future work.

For the tests of Cronbach's Alpha, multitrait scaling, factor analysis, we used the sample for whom we had complete data for both versions of the questionnaire (N=2737). For tests of discriminant validity where we examined the scale scores stratified by the number of comorbidities, we also based analysis on complete data for both questionnaire versions of the SF-36.

3. Results

Table 1 is the demographics of the sample. Over 90% were 65-99 years of age, 81% were white, 9% black and 5% Hispanic. 98% were male and about 72% were married. On average, subjects had more than 2 medical comorbidities and about 0.2 mental comorbidities. The demographic profile reflects for the most part the profile of veterans utilizing VA care.

Tables 2-4 give the Cronbach's Alpha statistics for the V/SF-36 and the MOS SF-36. For the V/SF-36, Cronbach's Alpha ranged from 0.86 for general health to 0.96 for role-physical, and for the MOS SF-36 from 0.85 (general health, social functioning and mental health) to 0.94 for physical functioning. No appreciable differences were found except for role physical and role emotional scales, where the V/SF-36 yielded consequential improvements over the MOS SF-36 (0.96 versus 0.91, for role physical and 0.95 versus 0.89 for role emotional, respectively). The correlations without overlap for the role-physical items ranged from 0.88 to 0.91 for the V/SF-36 and 0.76 to 0.82 for MOS SF-36. For the role-emotional items the correlations ranged from 0.91 to 0.94 for the V/SF-36, and 0.82 to 0.86 for MOS SF-36. The correlations without overlap were substantially higher for the V/SF-36 than the MOS SF-36 for these two scales indicating greater item convergent validity and internal consistency at the item level for the V/SF-36. This suggests that because of greater precision for the role items, the item-correlations for each concept are higher for the V/SF-36 version.

The Cronbach's Alpha for the Physical (PCS) and Mental Summaries (MCS) for the V/SF-36 and MOS SF-36 are:

	MOS SF-36	V/SF-36
PCS	.946	.956
MCS	.898	.946

Results suggest improvement in precision for the MCS summary of about 5% and 1% for the PCS summary.

Tables 5 and 6 portray the results of the multitrait scaling. For each scale, the correlations of its hypothesized items are shown with all 8 scales, including the hypothesized scale (shown in bold, corrected for overlap) and for the remaining 7 scales. For the role physical scale, the V/SF-36 yielded item-scale correlations without overlap with the hypothesized scale that was higher than the correlations with other scales. In all cases, the correlations were more than 2 standard errors higher than the other correlations, indicating that all were scaling successes for that scale. Similarly, the MOS SF-36 yielded all scaling successes for the role physical scale items. For the V/SF-36 and MOS SF-36, a similar pattern emerged for role emotional, with all scaling successes for both versions of the scale. Not surprisingly, the patterns of correlations in terms of scaling successes were similar for the other 6 scales for the V/SF-36 and MOS SF-36. Almost all correlations in the two instruments reflect scaling successes at the item level for each scale.

Table 7 reports the factor structures for the V/SF-36 and the MOS SF-36 for the total sample. Results indicate that the cumulative variance is about 3% higher for the V/SF-36 than the MOS version, 76% vs. 73%. This suggests greater explained variance, for the two factor model. This is because of greater precision in the V/SF-36 version. The pattern of the factor structures is similar for the two versions. Loadings indicate two factors, the first assessing physical health, and the second mental health. Communality estimates range from 0.65 to 0.89 for V/SF-36 and 0.67 to 0.81 for the MOS SF-36. The role physical communality was substantially higher for the V version and almost comparable for the role emotional for both versions.

Tables 8-16 gives the factor structures, separately by VA and HOS samples, for persons with specific chronic conditions (hypertension, low back pain, diabetes mellitus, chronic obstructive pulmonary disease, angina, congestive heart failure, heart attack, stroke, and depression). These conditions were chosen because of their prevalence in the VA and impact on health related quality of life. The factor structures reflect similar and almost comparable patterns of loadings for the two SF-36 versions. We do note that with the exception of diabetes and depression, all other conditions yielded slightly higher cumulative variance explained for the two factor structure for the V/SF-36 version than the MOS version. Diabetes and depression were comparable for the cumulative variance explained. Communality estimates were substantially higher for the role physical scale in the V version for all chronic conditions, while higher for role emotional for low back pain, angina, and depression. Higher estimates again reflect greater precision in terms of explained variability for that concept using a two factor solution.

Tables 17- 26 are the discriminant validity tests for each SF-36 scale and for the 2 summaries, comparing scale scores of the V and MOS versions among levels of physical and mental comorbidity. That is, for each SF-36 scale, means were estimated, separately for each version, HOS and VA, for varying levels of physical and mental comorbidity. Means are presented for 0 to 6 or more medical comorbidities and for 0 to 2 or more mental comorbidities (again derived from the ICD9-CM codes from the VA data sets).

We first focus on the role scales and the physical and mental summaries as they are distinctly different for the two SF-36 versions. Table 18 is the result of the role physical scale. The scores for the two versions indicate a significant monotonic relationship; those with zero comorbidity have the highest scores, while those with 6 or more comorbidities have the lowest scores. However, results indicate a much lower floor for the role-physical for the V/SF-36 than the MOS/SF-36 versions. The range of mean scores for the V version is from 47.05 for zero comorbidities to 17.62 for 6 or more comorbidities; and for the MOS version, the range of means is from 56.27 to 27.13. The floor of the V version is substantially lower than the MOS version. The ratio of the F statistics describing the monotonic trends is 11% more efficient for the V version using the number of medical comorbidities. For the number of mental comorbidities, role-physical scores indicate highly significant monotonic trends for the MOS and V versions. The range of mean scores for the V/SF-36 is from 36.54 to 22.82; and for the MOS/SF-36 version is 45.95 to 32.82. The ratio of the F statistics is 31% more efficient for the V version using the number of mental comorbidities and the floor is substantially lower for the V version. As expected, the differences in the monotonic trends for the two versions of the role physical scale are not significant for the number of medical and mental comorbidities. This suggests comparability in terms of the overall metrics for the two versions.

Table 23 shows the tests of discriminant validity for the role-emotional scale for the two versions. The range of levels for the role-emotional scale indicates a highly significant monotonic relationship for the number of medical and mental comorbidities for the V and MOS versions. For the number of medical co-morbidities, the scores for the V/SF-36 version range from 70.00 to 36.33 and for the MOS/SF-36 range from 73.76 to 50.05. For the number of mental comorbidities, the V version ranges from 60.16 to 26.93 and for the MOS version from 69.15 to 39.66. The floor for the V version as compared to the MOS version is substantially lower using the number of medical and mental comorbidities. We do note a significant F statistic for the difference between the V version and the MOS version for the number of medical comorbidities. This is likely the result of some differences in the linear trends at greater numbers of comorbidities. This may be partially due to smaller sample size groups and some imprecision in the estimates being compared at levels of four, five and six or more medical comorbidities. The differences in the monotonic trends of the two versions for the number of mental comorbidities are not significant. Importantly, as already noted the floor of the scale for six or more medical comorbidities is substantially lower for the V version than the MOS version. Of particular note the ratio of the F statistics describing the monotonic trends is 54% more efficient for the V version compared with the MOS version for this scale using the number of diagnosed medical comorbidities and about 5% less efficient for the mental comorbidities. Perhaps the slightly lower efficiency of this scale is a function of under reporting of mental diagnoses and consequently less reliability in this diagnostic indicator of mental health problems using the ICD-9 CM codes.

Table 25 displays the results for the physical summary score (PCS). The V and MOS versions display a significant monotonic trend across the number of medical comorbidities and approaches or are significant for the number of mental comorbidities. The difference between the two versions is not significant for both the number of medical and mental comorbidities, suggesting that the overall monotonic trend is not significantly different between the two versions. The floor of the V version for both medical and mental comorbidities is lower than the MOS version, at 6 or more and 3 or more medical and mental comorbidities, respectively. The relative efficiency of the MOS version is about 7% greater for the sum of the medical comorbidities than the V version, for the mental comorbidites it also favors the MOS version by about 56%. However, the fewer categories of the number of mental comorbidities, may not allow us to adequately discriminate the increase in precision of the V/SF-36 version.

Table 26 gives the discriminant validity of the mental summary scores using the MOS/SF-36 and the V/SF-36. Both the MOS and V versions display significant monotonic trends for the number of medical and mental conditions. The floor of the V version is lower than the MOS version for the number of medical and mental comorbidities. The difference between the two versions by level is significant for both the number of medical and mental comorbidities, suggesting differences in the monotonic relationship. The slopes between the two versions are different suggesting greater precision in the V version in terms of the ability to discriminate across the 7 levels. The V version is 44% more efficient for the medical comorbidities and about 11% more efficient for the mental comorbidities.

Results for the other six scales are displayed in tables 17, 19-22, and 24. For comparisons between the two versions, we note that the differences in these scales are not in their content or response choices but possibly in time lag between the administrations as well as the order of administration.. Of note the monotonic trends for the MOS/SF-36 and V/SF-36 are significant for the two versions for the medical and mental comorbidities.

4. Discussion and Conclusions

Results indicate important improvements to the precision and reliability of the V/SF-36 for the role scales (physical and mental) of 5% and 6%, respectively. These scales have a much higher alpha because they use a 5-point response scale instead of a 2-point response scale in the SF-36. The reliability of the physical and mental summaries also show an improvement of 1% and 5%, respectively. The item convergent validity was higher for the V version for the role physical and role emotional items an indication of increased precision. Multi-trait scaling suggested all scaling successes for the MOS and V versions. Factor analysis yielded comparable two factor structures overall and for the most part by selected chronic conditions. The factor structure yielded overall variances for the two factor structure that were greater for the V version than the MOS version, reflecting the greater precision in the role scales.

Discriminant validity of the role scales for the V version noted important increases to the efficiencies of the scale and also for PCS and MCS summaries. Further, the lowering of the floor was also marked for the scales and somewhat lower for the summaries. Interestingly, while there were no content changes to the stem or response choices of the items making up the six other scales, we note significant differences between the two versions for the number of medical comorbidities by the level of social functioning. We also note significant differences for the number of mental comorbidities by the level of mental functioning. Because the questionnaires were administered to the same subjects we suggest that there are several possible reasons for these differences. The first is the order of administration of the questionnaires. There is the possibility of a survey effect. While the MOS version was fielded first from Cohort 2, a fraction of patients returned the MOS questionnaire after they returned the VA questionnaire. Order of administration will be considered in future work. Related to this is the lag between the administrations of the V and MOS versions, which also will be considered. This contributed to different scores among the levels of comorbidities when comparing the two versions. Second, the veterans may have been sensitized to the VA questionnaire, and potential concern over losing disability coverage may have driven the scores lower when compared to the HOS survey where that concern is not of much consequence (Medicare coverage is mandated, while VA coverage may be based upon service connected disability). Third, the mode of administration may also be a factor to consider. The VA survey involved mail out administrations (V version); however the HOS survey used a mixed approach, where a fraction of respondents were given a telephone administration as a second protocol to the non-responders of the mail out waves. As we have previously documented in the literature, there is a contextual effect related to mode of administration. Telephone administration yields higher scores than a mail out approach (Jones et al. 2000). Fourth, formatting differences between the two versions may have also influenced differences between results. The questions in the V version make use of both vertical and horizontal lines to make responses to questions more easily identified. More space is also used between items and the response choices in the V version. Although discriminant validity appears to be improved for the mental health scale this is likely not a function of changes in the metric but perhaps a combination of the factors noted above.

For the mental component summary (MCS), the highly significant finding of differences in the level of mental functioning by the number of medical conditions with improved efficiencies for the V/SF-36 version, is likely a consequence of improved precision to the role-emotional scale.

The improvements in the precision of the role scales for the V version are clearly important. The comparability on the other measurement properties between the two versions for multi-trait scaling and scale level factor analysis suggests that for the most part the other 6 scales behave in fairly comparable ways. Future studies will control for the sequence of administration of the questionnaires and lag between them that could have affected the results.

Future work will consider other measures of capturing the discriminant validity of the scales including measures of disease severity for specific chronic conditions, such as diabetes, chronic lung disease, osteoarthritis, chronic low back pain and major depression. Clinical measures available on patients through the VA data merge will be evaluated and used as external measures of validity in association with the V/SF-36 and MOS SF-36. We will also consider the change scores and two stage models of change for cohort 2 and compare the responsiveness to change of the V/SF-36 with the MOS version.

The V/SF-36 is an important assessment tool alternative to the MOS SF-36 given improvements to the precision of the V version for the role scales and the component summaries and comparability in terms of multitrait scaling and factor analysis. The results provide support for the greater reliability and validity of the V/SF-36 over the MOS SF-36 versions; they also provide evidence that the psychometric properties and measurement characteristics of the V/SF-36 are at least comparable to the MOS version, and in fact better for selected scales and component summaries. Thus, our results lend support that the two versions of the SF-36 can both be used to conduct future system (Medicare versus VA) comparison studies.

5. References

Hays RD and Stewart AL. The structure of self-reported health in chronic disease patients. Psychological Assessment: A Journal of Consulting and Clinical Psychology 1990;2:22-30.

Kazis LE, Wilson N. Health Status of Veterans: Physical and Mental Component Summary Scores (SF-36V): 1996 National Survey of Ambulatory Care Patients, Executive Report. Office of Performance and Quality, and Health Assessment Project, Health Services Research and Development Service. Washington D.C. and Bedford, Massachusetts, March 1997.

Kazis LE. et. al. (a) Health Status of Veterans: Physical and Mental Component Summary Scores (SF-12V). 1997 National Survey of Ambulatory Care Patients, Executive Report. Office of Performance and Quality, Health Assessment Project HSR&D Field Program, VHA National Customer Feedback Center, Washington, D.C., Bedford and West Roxbury, Massachusetts, April 1998.

Kazis LE. et. al. (b) Health Status and Outcomes of Veterans: Physical and Mental Component Summary Scores (SF-36V). 1998 National Survey of Ambulatory Care Patients, Mid-Year Executive Report. Office of Performance and Quality, Health Assessment Project, HSR&D Field Program, Washington, D.C. and Bedford, Massachusetts, July 1998.

Kazis LE, Miller DR, Clark J, Skinner K, Lee A, Rogers W, Spiro III. A, Payne S, Fincke G, Selim A, Linzer M. Health related quality of life in patients served by the Department of Veterans Affairs: Results from the Veterans Health Study. Arch Intern Med. 1998;158:626-632.

Kazis L, Ren XS, Lee A, et al. Health status in VA patients: Results from the Veterans Health Study. American Journal of Medical Quality 1999; 14(1):28-38.

Kazis LE. The veterans SF-36 health status questionnaire: Development and application in the Veterans Health Administration. Medical Outcomes Trust Monitor 2000; 5 (1).

Kazis L, Miller D, Clark J, Skinner K, Lee A, Ren XS, et al. Improving the response choices on the SF-36 role functioning scales: Results form the Veterans Health Study. Medical Care, Supplement 2003, (in press).

NCQA (1998). HEDIS® 3.0/1998, Volume 6: Health of Seniors Manual. Washington, DC: National Committee for Quality Assurance.

Perlin J and Kazis LE, Skinner K, Ren XS, Lee A, Rogers W, Spiro A, Selim A, Miller D. Health Status and Outcomes of Veterans: Physical and Mental Component Summary Scores, Veterans SF-36, 1999 Large Health Survey of Veteran Enrollees, Executive Report. May 2000. Department of Veterans Affairs, Veterans Health Administration, Office of Quality and Performance, Washington, D.C.

Selim AJ, Fincke G, Ren XS et al. Comorbidity assessments based on patient report: Results from the Veterans Health Study. Med Care Supplement (2003).

Ware JE Jr, Sherbourne CD. The MOS 36-item short form health survey (SF-36) I. conceptual framework and item selection. Med Care 1992;30:473-83.

Ware JE Jr. with Snow KK, Kosinski M, Gandek B. SF-36 Health Survey: Manual and Interpretation Guide. The Health Institute, New England Medical Center, Boston, MA. 1993.

Ware JE Jr., with Kosinski M, Keller SD. SF-36 Physical and Mental Health Summary Scales: A User’s Manual. Health Assessment Lab, New England Medical Center, Boston, MA. December 1994.