

Interpreting SF-36 Summary Health Measures: A Response

John E. Ware, Jr.^{abcd ‡}, PhD and Mark Kosinski^{a*}, MA,

^a Quality Metric, Inc.
Lincoln, RI

^b Health Assessment Lab
Boston, MA

^c Harvard School of Public Health
Boston, MA

^d Tufts University School of Medicine
Boston, MA

Address all communications to: John E. Ware, Jr., PhD, QualityMetric, Inc. 640 George Washington Hwy, Suite 201, Lincoln, RI, Phone: 401-334-8800, x242, Fax: 401-334-8801, E-Mail: jware@qmetric.com

Key words: Factor analysis, Health-related quality of life, physical and mental health status, Medical Outcomes Study (MOS), PCS and MCS summary health measures, Questionnaires, SF-36 Health Survey

Introduction

The questions raised by Taft et al. [1] (Taft) regarding the “accuracy” of *SF-36* physical (PCS) and mental (MCS) component summary scores are important and we thank the editors of the *Journal* for this opportunity to comment on their conclusions. In response, we test the assumptions underlying Taft’s hypotheses and describe the internal “workings” of the *SF-36* items, scales and summary measures. Finally, we illustrate how analyses of external criteria can be used to test the validity of extreme PCS and MCS scores.

PCS and MCS were developed from studies of the physical and mental higher-order factors identified during the early years after the development of the *SF-36* [2,3]. These studies were conducted to evaluate the psychometric properties of the *SF-36* including construct validity and the factor content of each scale. When we realized that these summary measures capture more than 80% of the reliable variance in the eight subscales, we pursued a second goal, which was to develop psychometrically-based summary measures that simplify the analysis and interpretation of the *SF-36*. For both purposes, we struggled with numerous methodological considerations. When the two components are rotated to simple structure should they remain orthogonal as they were when extracted or rotated to be oblique (correlated)? If oblique, how large should the correlation be? For the first goal, it is clear that orthogonal health components are most useful [4,3,5]. For the second goal, the discussion goes on. One way to think about the issues is in terms of the following questions: How much physical health should be in a mental health summary measure? How much mental health should be in a summary measure of physical health? In very straightforward ways, the answers determine the validity and interpretation of their scores. Because the *SF-36* summary measures are widely used for a variety of purposes, we believe that these questions should be addressed in terms of the total weight of the evidence and not on the basis of results from any one study.

Concerns about extreme PCS and MCS scores were salient from the outset because we considered adoption of the same 0-100 (lowest to highest, respectively) score transformation that we

had adopted for the subscales in the *SF-36* profile. The 0-100 transformation required that we define both extremes and that we decide how to handle any scores that appeared questionable (e.g., scores outside the theoretical range). At that time we estimated summary scores for those scoring zero and 100 across all subscales and considered truncating scores for those outside the range of PCS and MCS summary measures defined by those extremes. After much discussion and empirical study, we concluded that theory was weak with regard to these extreme scores and that their information value warranted further study. If we had “rushed to judgment” and truncated these extreme scores at the outset, their usefulness may not have been evaluated by others in the field.

We also adopted norm-based scoring (NBS) for the PCS and MCS because the “ceiling” (highest possible) and “floor” (lowest) scores were destined to be raised and lowered, respectively, to meet the increasing demands of the growing field of outcomes research. Without these improvements, too many people were being “lumped” at the ceiling and at the floor. NBS, which estimates scores in standard units as deviations from the average (rather than from the extremes), makes it possible to extend the range without changing the interpretation of scores in between. If we had continued with our original 0-100 *SF-36* transformation, every improvement in the range covered in either direction would have changed the meaning and interpretation of scores in between. It should be noted that, because NBS was implemented using a simple *linear* transformation, no “statistical” conclusions (correlations, probabilities, F-ratios, etc.) were changed [5].

We observed early on that orthogonal (uncorrelated) scores for the two principal health components best discriminate between physical and mental health outcomes. However, we accepted from the beginning that these components would not be as valid as the best subscale, particularly when differences are concentrated in one subscale. We offered numerous examples of this potential shortcoming of the PCS and MCS summary measures and its implications [4,3]. Because of the potential for information loss with summary health measures, we encouraged those who use them to interpret their results in parallel with the profile of

SF-36 subscales and we developed computer software to facilitate such comparisons. We continue to study the tradeoffs between the simplicity of two summary measures and the richness of a profile of health outcomes and we continue to evaluate different approaches to scoring them and we encourage others to address these issues and to report their results. The number of peer-reviewed articles about the *SF-36* is growing. We are aware of 120 publications about PCS and MCS from 1994 through 2000; most were published in the past two years. We leave it to others to judge whether Taft is accurate in characterizing this number of publications as “relatively few empirical studies.”

Scoring Principal Component Summary Measures

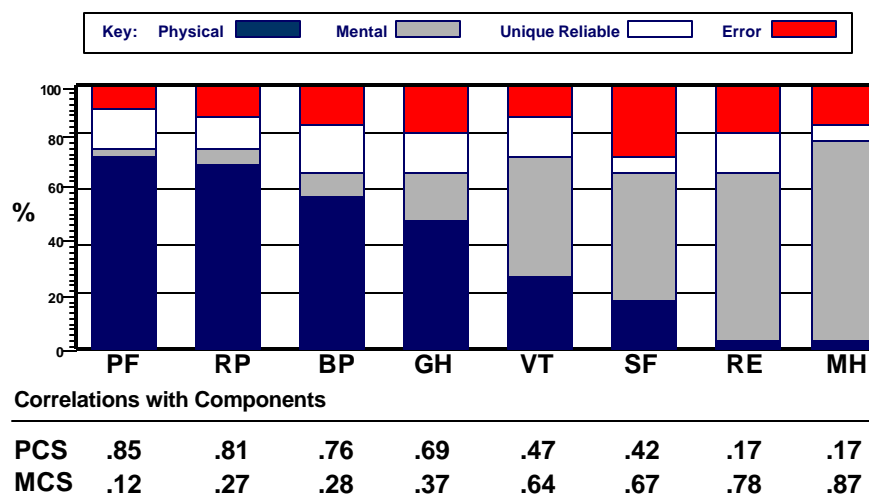
Taft is simply incorrect in stating that because nine positive coefficients and seven negative coefficients are used in scoring the PCS and MCS, these summary measures define *opposing* (physical versus mental) components of health. Specifically, Taft concludes, after summarizing the steps involved in their scoring, that “implicit” in this approach is the assumption that “these two components of health are in effect opposed.” To the contrary, the PCS and MCS are not negatively correlated and *all* eight subscales have correlated *positively* with both components in all countries studied to date. As shown in Figure 1, the positive correlations between the subscales and each component are ordered in terms of their magnitude from left to right for the highest (PF) to lowest (MH) correlations with the physical component and from right to left in terms of the highest (MH) to the lowest (PF) correlations with the mental component.

Perhaps, the apparent confusion is a consequence of Taft’s attempt to interpret PCS and MCS using factor score coefficients. Psychometric texts recommend interpretation of components in terms of their *correlations* with the variables being analyzed. For principal components, it is the subscale-component correlations and not the scoring

coefficients that are the key to interpretation. In our peer-reviewed articles and manuals, the scoring coefficients are not even mentioned for the purpose of interpretation. For oblique factors, both scoring coefficients and subscale-factor correlations are difficult to interpret because of the substantial overlap between oblique factors. For example, according to their squared correlations with the subscales, oblique physical and mental health factors explain about 125% of the reliable variance and about 150% of the common variance in the eight *SF-36* subscales, i.e., much more than is possible. The extra variance, of course, is that added twice to the scores for the two oblique factors.

What is the factor content of the *SF-36* subscales and how do positive and negative component weights improve their validity in discriminating between physical and mental health outcomes? Figure 1 below shows graphically the factor content of *SF-36* subscales and their component correlations [5]. The total variance in each *SF-36* subscale is divided into four parts accounted for by: (1) the physical (PCS) component of health (black bars), (2) the mental (MCS) component of health (gray bars), (3) the unique reliable variance, i.e., reproducible but not accounted for by either component (white bars), and the error, which is 1 minus the reliability (cross-hatched bars at the top). It should be noted that every subscale manifests a different pattern and that all subscales are measured with some error, usually 10-20% or less. PF and MH are the “purest” measures of the physical and mental components, respectively. The most complicated in terms of factor content, some would say the most confounded, subscales are the middle four (BP, GH, VT and SF). (This should not be surprising; SF items, for example, ask about both physical and mental health status.) When these scales are aggregated in computing a summary score, substantial confounding is introduced because each of these subscales adds information about more than one component of health. Negative coefficients remove the health scores that, otherwise, would be counted twice.

Factor Content of SF-36 Subscales, General U.S. Population



Source: Ware and Kosinski, 2001

Figure 1

Principal component scores provide a proven solution to the confounding problem. The positive and negative coefficients involved are fundamental to the scoring of variables (subscales) that have complicated factor content, i.e., measure two or more components. They are not unique to orthogonal components. For example, in scoring PCS, scores are aggregated using positive weights for the first five subscales (PF through VT), which contribute the most information about the physical component of health (the black parts in Figure 1). However, these subscales also bring with them substantial variance due to the second component (mental health; the gray parts in Figure 1). For a given individual, if the scores for the second part (a different health outcome) are above the mean, they must be subtracted back out to keep from inflating the score estimate for the physical component. Likewise, if the scores for the second part are below the mean they must be added back, to keep from introducing downward bias into the estimate of the physical component score. The same holds true for the four scales on the right (VT through MH) in Figure 1, which are weighted positively in estimating the mental component. When they are added to the MCS score, substantial information

about a different health outcome (physical health, the black bars) is also added to the estimate of the mental health component score. To correct for the confounding of physical and mental health, negative coefficients for some subscales subtract back out the unwanted variance.

Proponents of orthogonal and oblique scoring algorithms for summary health measures differ in terms of the amount of correction they make or, in other words, the amount of confounding or overlap they are comfortable with when scoring and interpreting summary scores. The PCS and MCS scores are orthogonal. PCS is an aggregation of the black bars (no gray bars) in Figure 1 and MCS is an aggregation of the gray bars (no black bars). By minimizing their overlap (confounding), their validity in measuring one (but not the other) component of health outcomes is maximized. In sharp contrast, oblique higher order factors derived from the SF-36 have substantial overlap as evidenced by their substantial or high inter-factor correlations (about 0.50 to 0.70 in published studies we are aware of). The practical implications are largely unknown. With the exception of our own recent presentation [6], we are not aware of any published studies that have used oblique SF-36

factors to estimate disease burden or treatment benefits.

Finally, with regard to scoring, we were unable to follow Taft's discussion of how the contributions of the SF-36 subscales to the PCS and MCS components varies independent of their weights (scoring coefficients). The principal component scoring method standardizes subscale variances prior to applying the weights. The same holds true for the scoring of oblique factors. SF-36 subscales are no longer in their original (0-100) units, they are z-scores, when they are weighted and aggregated to score PCS and MCS [5].

The Meaning of High and Low Scores – the inner workings of the SF-36

Taft offers hypotheses about PCS and MCS scores outside the "expected" range (outside of 20-58 for PCS and 17-62 for MCS). Based on untested assumptions regarding the pattern of subscale scores underlying such summary scores, Taft argues that they "are a product of negatively weighted subscales" used in scoring PCS and MCS. (Note that for PCS three subscales are weighted negatively as are four for MCS [4]. Taft questions the "accuracy" of these extreme scores based on the assumption, for example, that a PCS score below 20 or MCS below 17 occurs only when someone scores zero on the subscales (0-100 scoring) in a physical or mental cluster. (Throughout this response, we will refer to PF, RP, BP, and GH as the cluster of physical subscales; the mental cluster includes MH, RE, SF & VT: their abbreviations are documented in Figure 1.) If those scoring below 20 on PCS score zero on the physical cluster of subscales, the only way to score lower on PCS would be to score higher (better) on one or more of the three scales that are weighted negatively in scoring PCS, e.g., the Mental Health (MH) subscale. If Taft is correct, it is reasonable to assume that PCS scores lower than 20 would be invalid because they would only reflect an improvement in mental health rather than a worsening in physical health. Likewise, if PCS scores above 58 are earned only by those scoring 100 on subscales in the physical cluster, those PCS scores would be invalid because they would reflect worse mental health rather than better physical health. If Taft is correct, the same logic argues for the invalidity of MCS scores below and above 17 and 62, respectively.

Taft's logic appears to be reasonable if the underlying assumptions are correct with regard to the patterns of subscale scores that account for summary measures at the extremes. Surprisingly, Taft did not test these assumptions. In the combined MOS and general US population data sets (N = 6,742), both available for public use, scores outside Taft's "expected" range have been observed, but only rarely at the lower levels (less than 1-2%); they occurred more often (up to about 10-15%) at the higher levels). Only one person (out of 102) who scored below the "expected" range did so in the manner hypothesized by Taft, i.e., by scoring zero on subscales in the physical cluster. Among the much larger sample of respondents who scored higher than the "expected range (N=1,119), a small minority did so by scoring 100 on the subscales in the physical and mental health clusters, i.e., only 168 of 1,119 (15%) for PCS and only 74 of 319 (23%) for MCS. Only four (out of 59, 6.8%) of those scoring outside the low range of MCS did so by scoring zero on subscales in the mental health cluster.

If not the invalid patterns hypothesized by Taft (about 15% of respondents), what patterns of responses accounted for scores outside the range (about 85% of respondents)? To answer this question, we examined results from published cross-tabulations of responses to SF-36 items for those scoring at the extremes of the PCS and MCS score ranges (see Tables 9.2 & 9.3 in Ref [4]), which coincidentally include categories of scores outside of Taft's limits. Consistent with the above descriptive statistics, no one scoring at or below 20, the lowest PCS score "expected" by Taft, earned scores of zero on the subscales in the physical cluster. For example, 18% of those in the 8-29 PCS range could walk 100 yards (PF subscale > 0), less than half reported very severe pain (BP subscale > 0), nearly one-fourth rate their health as "excellent to good" as opposed to "fair or poor" (GH subscale > 0). At the other extreme of the PCS range, the cross-tabulations also disprove Taft's basic assumption. For example, 20% reported limitations in vigorous activities (PF subscale < 100) and less than half rated their health as "excellent" (GH subscale < 100); interestingly, those at the highest levels of physical health do not necessarily have more energy (VT subscale). For MCS, we observed a similar pattern of item responses contradicting Taft's assumptions. For example, 10% or more of those

above 62 on MCS are “happy” less than all or most of the time (MH subscale < 100) and more than a third do not report “a lot of energy” less often than all or most of the time (VT subscale < 100) as required to score 100 on these subscales.

Thus, in contrast to Taft’s hypotheses, scores for subscales in the physical health cluster determine variation in scores for most respondents scoring at the “ceiling” for PCS and subscales in the mental health cluster determine variation in scores for most respondents scoring at the “ceiling” for MCS.

The Meaning of High and Low Scores – external criteria

Criterion evidence (including both concurrent and predictive tests) in the SF-36 user’s manuals is also relevant to the issue of whether extreme PCS and MCS scores are valid [4]. What are the hypotheses underlying these tests? Taft hypothesizes that scores increase above the “expected” range not because health is getting better but because the other component of health is actually worse. Below the expected range, scores go down on one component because scores for the other are actually getting better. In a nutshell, Taft hypothesizes that the well-documented monotonic trends in PCS and MCS scores shift to nonmonotonic trends outside the “expected” range.

What do criterion tests of validity show? They confirm the monotonic trends, which are the basis for scoring and interpreting the PCS and MCS. For example, in comparison with scores within Taft’s “expected” range, those scoring extremely high on the MCS (e.g., 60-74) are more than one-third less likely to screen positive for depression and are about one-third more likely to report being satisfied with their life in comparison with those in the 55-59 and 50-54 MCS categories. The rest of the monotonic trend for this and other criteria is documented elsewhere (see Table 9.10) [4]. If MCS scores higher than 62 are an artifact of lower physical health scores (weighted positively in scoring MCS), as argued by Taft, we would expect the opposite pattern of results from these “criterion” tests of validity. Taft also questioned the validity of very low MCS scores. Published criterion tests support the validity of MCS scores in the extreme low range. For example, those with MCS scores in the 3-29 range were 50% more likely to have a

diagnosis of clinical depression and 40% more likely to receive mental health specialty care, in comparison with those with MCS scores in the 30-34 range. These percentages decrease monotonically as MCS scores increase.

Are PCS scores lower than 20 valid?

Contrary to Taft’s hypothesis, results published for chronically-ill adults participating in the Medical Outcomes Study (MOS) [4], see Tables 9.4 – Table 9.8] show that patients scoring in extremely low PCS categories (8-24, 8-29 and 8-34, across various criteria) had significantly higher rates of job disability, job loss within one year, subsequent hospitalizations, greater disease burden, and greater likelihood of death within five years, in comparison with the next highest PCS score group. In support of the validity of PCS scores, all trends in results for these criteria were monotonic. In analyses of 5-year mortality, for example, the rates for five PCS groups were (see Table 9.8 in Ref [4]): 21.5% for PCS = 8-24; 15.1% for PCS = 25-34 (an increase in mortality of more than 40%); 6.2% for PCS = 35-44; 4.7% for PCS 45-54, and 1.8% for PCS 55-72. These results do not support Taft’s hypotheses regarding the invalidity of PCS scores at either extreme (below 20 or above 58). However, because the lowest category above (PCS = 8-24) was divided nearly equally between those scoring above and below 20, i.e., within and outside the range expected by Taft, the results above are somewhat ambiguous. Hence, we sought additional evidence.

Although PCS scores below 20 are very rare in the U.S. population, as they appear to be in Sweden, large ongoing studies that are using the PCS to predict 2-year mortality in the U.S. Medicare population (ages 65 and older) permit more definitive tests of hypotheses regarding the validity of very extreme PCS scores. For example, in the first cohort of the Medicare Health Outcomes Survey (HOS) (N = 172,314), less than 3/10ths of one percent, scored below 20 on the PCS. Still, because of the size of the HOS, that sample (N = 432) proved to be large enough to perform predictive tests of the validity of extremely low PCS scores. Preliminary results indicate that the 2-year mortality rate for HOS participants with baseline PCS scores below 20 was about one-fourth higher than the rate for those scoring 21 to 30 (21.4% versus 17.3%, respectively).

We observed analogous results supporting the validity of extremely high PCS scores in the

MOS and HOS studies. Although the mortality rates are much lower for all groups with very high PCS scores, as expected, the differences between groups were significant and consistent with hypotheses for a valid physical health measure. For example, the 5-year mortality rate for those scoring 58 and higher on the PCS (N = 19,161), replicate the MOS results. For example, 2-year mortality rates increased from 2.2% for those scoring above 58 and higher on PCS (N = 28,539) to 3.0% for those scoring 51-55 on PCS (N = 29,977) and increased to 4.2% for those scoring in the 46-50 range on PCS (N = 23,086). Thus, those scoring 58 and higher on PCS appear to be in better physical health.

The results from criterion and predictive tests summarized above support the validity of extreme PCS and MCS scores, including those in the extreme ranges where Taft concluded that these scores are invalid. Negative weights used in scoring PCS and MCS do not appear to render their scores at the extremes invalid, as hypothesized by Taft.

Cross-sectional Comparisons of Profiles and Summary Measures in Sweden

To illustrate the practical implications of hypothesized problems with the PCS and MCS scoring algorithms, Taft compared scores for younger and older adults in Sweden (see Figure 4 in [1]). In the figure, for younger adults, it appears that scores for the cluster of mental health subscales are above the norm for younger adults whereas the MCS score is below the norm. Certainly, this would be a cause for concern and the figure led Taft to conclude that the “MCS score falsely indicates poorer than average mental health” for younger adults in Sweden. We were unable to replicate the discrepancy apparent in Taft’s Figure 4 in analyses of data from Sweden and turned to SF-36 public use files for the U.S. population to see if we could replicate Taft’s finding among younger and older age groups in the U.S. We found no noteworthy discrepancies between MCS scores and scores for the cluster of mental health subscales. We focused our MCS comparisons on the MH subscale in both countries because that subscale is the “purest” mental health measure in the SF-36. MH is scored using the Method of Summated Ratings, a simple sum without weights, and is not subject to the criticisms Taft has directed at the MCS. Finally, MH is the most extensively validated of the SF-36

subscales and the focus of more than 250 published studies of emotional disorders.

We observed agreement between the MCS and MH subscale scores within one point, on average, for both younger and older age groups in the U.S. (profiles and scores available from the authors upon request). Because Figure 4 does not document the actual scores on which it was based, we also re-analyzed data for younger (< 35 years) and older (65+ years) adults published in the Swedish SF-36 manual [6] in an attempt to replicate the “apparent” discrepancy between MCS and MH in Sweden. We used standard NBS algorithms and SF-36 component score coefficients [5]; however, we substituted Swedish means and standard deviations (from [6]) for the eight subscales. Results from the re-analysis of data from Sweden appeared to be the same as results from the U.S. Specifically, the MCS summary scores and MH subscale scores were within one point of each other, on average, for both younger and older age groups. When results for both countries were plotted using standardized SF-36 scoring software and reports, the reason for the “discrepancy” visually apparent in Taft’s Figure 4 became clear. Taft’s Figure 4 was not drawn to scale.

When drawn to scale (using NBS with a mean of 50 and SD=10), the pattern of results for profiles and PCS and MCS summary measures for younger adults is virtually identical in Sweden and in the US. For those under age 35 in Sweden, we observed scores above their respective population norms for the four subscales in the physical health cluster (PF, RP BP and GH) and scores at the average (+/- 1 point) for the four scales in the mental health cluster (MH, RE, SF & VT). As expected from these subscale profiles, the average PCS score for younger adults was above the norm (PCS = 53.1) and at the norm for MCS (MCS = 49.6) in Sweden. In the US, a nearly identical pattern of results was observed (all subscales in the physical health cluster above the norm, subscales in the mental health cluster at the norm and PCS > MCS).

We suggest that Taft over-interpreted the results from Sweden. Among younger adults, a difference in MCS scores of only 4/10ths of a point (49.6 as opposed to 50.0) near the middle of the scale range (as opposed to a range indicating severe emotional distress) does not justify the conclusion that the MCS “falsely indicates poorer than average mental health” among younger Swedish adults. The

same logic applies to the conclusion of “significantly better mental health” for older versus younger adults in Sweden based on MCS scores (53.90 versus 52.91, respectively). Incidentally, the MH subscale scores for younger and older adults in Sweden were also within one point of each other.

We take this opportunity to discourage interpretations of differences in standardized NBS scores (SF-36 subscales and summary measures), close to the average, that amount to one point or less, pending evidence that they are clinically, economically and socially relevant. We are unaware of any such evidence. Finally, we have chosen to ignore the insignificant results based on small samples as reported by Taft. We don't know how small they were at the extremes of the scores in some of the Taft analyses because sample sizes were not reported. We appreciate Taft's cautionary note in this regard. From our aborted attempts to replicate them in the US, some samples were probably very small. For example, in the most recent US general population SF-36 norming survey there were only 11 people (a fraction of 1%) in one of the categories (PCS <18) for which Taft computed correlations between components. Regardless of sample size, it is not obvious what the cross-sectional differences in correlations should be between valid measures of physical functioning and well-being. For example, from the other extreme, where samples are larger, if those who are most physically active in terms of vigorous activities tend to report lower vitality scores (e.g., report feeling tired more often) is that a problem of measurement validity?

Final Comments

We thank Taft for stimulating our thoughts and hope that our comments help to increase understanding of the logic and methods underlying the principal components method we have used in deriving and scoring the SF-36 PCS and MCS summary health measures. As noted above, we find little or no support for Taft's hypotheses about the interpretation of scores for the PCS and MCS summary measures. However, we encourage replications by others before any conclusions are generalized.

Examples of how others have interpreted the PCS and MCS are available in at least 120

publications using those summary measures (we have listed the complete citations for these studies on the SF-36 community website at <http://www.sf-36.com/news/pcsmcscitations.rtf>). However, we need not limit ourselves to these studies in our evaluations of the performance of the SF-36 summary measures in relation to the profile of eight subscales. Because the PCS and MCS are simply weighted aggregations of scores for the eight SF-36 subscales, more than 2,000 publications reporting SF-36 profiles are currently being re-scored and compared using SF-36 NBS software. We have selected studies directly relevant to Taft's hypotheses and we have computed standardized profiles as well as orthogonal and oblique summary measures for these studies. Presentation of the resulting graphs and discussions of results from these studies are beyond the scope of this response. We present both on the SF-36 community website (at www.sf-36.com/nbs). As noted there, to date we find no support for Taft's hypotheses in the results from these studies.

Surprisingly, in their advocacy of oblique factors, none of the critics of orthogonal components cited by Taft mention any of the tradeoffs that might be involved in scoring SF-36 summary measures on the basis of oblique factors. We take this opportunity to point out: (1) like orthogonal components, oblique factors are less responsive when outcomes are concentrated in one subscale; (2) oblique factors also require negative scoring weights (for 5 of the 8 subscales); (3) oblique factors dilute the distinction between physical and mental health outcomes; (4) when scores are very low for one component of health, oblique factors often underestimate scores for the other component; and (5) the greater the correlation between health factors the more dependent high and low scores on one are on the same pattern for the other. In the case of oblique factors, for example, the question becomes: Why should the highest physical functioning score require that someone also be happy all of the time?

As we have recommended in all of our publications about PCS and MCS, one of the best defenses against inappropriate conclusions based on the summary measures is the thorough comparison with results based on the eight SF-36 subscales. This logic also works well in the other direction. Unexpected differences observed in one subscale (often the RP or RE subscale), can be scrutinized in terms of whether they are substantiated by the more

comprehensive and less coarse PCS or MCS scores. These comparisons can be confusing because SF-36 subscales have been reported in their original 0-100 metrics and the PCS and MCS have been reported using NBS (mean = 50, SD = 10) in nearly all studies published to date. Perhaps, the best way to compare them is to compare standardized scores for both the profile and PCS and MCS. To make such comparisons easier, we have made the NBS utilities used in scoring subscales and summary measures for all SF-36 manuals available on the Internet (www.sf-36.com/nbs). Using this scoring utility software, SF-36 subscale scores (0-100) are transformed to have the same mean and standard deviation (50 and 10, respectively) as PCS and MCS. Further, the SF-36 scoring utility instantly prints out profiles and summary scores along with graphs that make results directly comparable for subscales and summary measures. This scoring utility is available for use with the SF-36 in Sweden and the U.S. on the Internet at www.sf-36.com/nbs. To facilitate comparisons between results based on orthogonal (PCS and MCS) and oblique (correlated physical and mental factor scores), we have added estimates

of oblique factor scores to the SF-36 scoring utilities and we have added both orthogonal and oblique summary measures to the graphs included in the output. We hope that others will use these scoring utilities, as we have, to compare their results and that the utilities prove to be useful in deciding which approach best communicates SF-36 results.

Acknowledgements

We gratefully acknowledge Chris Dewey at QualityMetric for his assistance in programming the software for norm-based scoring in Sweden and the U.S.; Martha Bayliss at QualityMetric for sharing results from her ongoing synthesis of the literature on treatment outcomes based on the SF-36 and Justin Sinclair and Barbara Gandek at the Health Assessment Lab (HAL) for providing preliminary estimates of mortality rates from the Medicare Health Outcomes Survey (HOS). SF-36 ® is a registered trademark of the Medical Outcomes Trust (MOT).

References

1. Taft C, Karisson J and Sullivan M. Do SF-36 summary component scores accurately summarize subscale scores? *Quality of Life Research*, 2001, in press.
2. McHorney CA, Ware JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36®): II psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993; 31(3):247-63.
3. Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36® health profiles and summary measures: summary of results from the Medical Outcomes Study. *Med Care* 1995; 33(Suppl. 4): AS264-AS279.
4. Ware JE and Kosinski M. *SF-36® Physical and Mental Health Summary Scales: A Manual for Users of Version 1. 2nd Edition*. Lincoln, RI: QualityMetric, Inc., 2001.
5. Ware JE, Kosinski M, Dewey JE. *How to Score Version 2 of the SF-36® Health Survey (Standard & Acute Forms), 2nd Edition*. Lincoln, RI: QualityMetric, Inc., 2001.
6. Sullivan M, Karlsson J, Ware JE. *SF-36 Halsöenkät: Svensk manual och tolkingsguide (Swedish manual and interpretation guide)*, Göteborg: Health Care Research Unit, Gothenburg University, 1994.
7. Ware JE, Kosinski M, Perfetto EM and Bjorner J. Comparison of treatment outcomes estimated using oblique & orthogonal physical & mental health summary scores: Results from 42 randomized trials using the SF-36 Health Survey. *Quality of Life Research*, 1999, 8 (7), p. 654.