Interpreting SF-36® Summary Health Measures: A Response

Supplemental Documentation

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: SF-36 health survey, health-related quality of life, factor analysis, PCS and MCS summary health measures, Medical Outcomes Study (MOS), health status, questionnaires

Acknowledgements: We gratefully acknowledge Chris Dewey at QualityMetric for his assistance in programming the software for norm-based scoring in Norway, 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).

We present here supplemental information that could not be published in *QOLR* due to space constraints. The *QOLR* editors were very generous in the space allowed in presenting our response to the questions raised by Taft, Karlsson and Sullivan (Taft). However, some important results of our literature review and re-analyses of published data had to be left out. We present that information here. First, we address the issue of whether summary SF-36 health measures based on oblique factors provide a solution to the potential problems with the scoring of summary health measures. **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.** Second, we look to the literature for evidence in support of Taft's hypotheses. In this supplement, we summarize published results from four studies and re-analyze their results to permit comparisons between orthogonal and oblique algorithms for scoring SF-36 summary physical and mental health measures.

Do Oblique Factors Provide a Solution?

About a handful of investigators have argued that summary physical and mental health measures based on oblique (correlated) factors prevent the "potential problems" caused by the negative factor score coefficients used in estimating orthogonal (uncorrelated) health components, particularly when differences in scores for these components are very large. Accordingly, Taft calls for revisions in the scoring of physical and mental health summary measures based on the SF-36. Unfortunately, Taft does not suggest an alternative. Others, cited by Taft, focus their criticisms of PCS and

MCS on the negative coefficients required to score PCS and MCS using the principal components method. For example, Simon and his colleagues (1998) argue that the potential for bias in component scores resulting from their use of negative scoring coefficients is greatest when one component differs by an extreme amount. They recommend correlated physical and mental summary measures based on an *oblique* rotation as the solution to this problem.

We compliment Simon and his colleagues for their thorough documentation of their methods and results. Their table with effect sizes for SF-36 subscales and summary measures is very useful because it standardizes changes in scores so that they can be meaningfully compared. We take this opportunity to point out, however, that *oblique* physical and mental health factors do not eliminate negative scoring coefficients for the SF-36 subscales. To the contrary, even when the degree of inter-factor correlation is substantial (e.g., about 0.55 using a Promax oblique solution), five of the seven scoring coefficients (that are negative in the algorithms for orthogonal components) are also negative in the oblique solution.

The concerns of Simon and his colleagues were expressed in an article reporting a study of depressed patients treated in primary care settings. From this study Taft concluded that "impaired physical health indicated in profile scores was not reflected in the PCS score at baseline." In fact, the extent of the agreement between the Physical Component Summary (PCS) and the eight *SF-36* subscales at baseline varied markedly in that study depending on which subscale was compared with the PCS. Completely inconsistent with Taft's conclusion, the average scores for the Physical Functioning (PF) subscale (the purest physical health subscale) and the PCS were both within about one

point of the population norm of 50 at baseline (49.4 and 51.3, respectively, for PF and PCS, standardized using norm-based scoring). More consistent with Taft's conclusion, averages for the three other subscales (RP, BP & GH) contributing most to the PCS summary were below the norm (43.2 to 45.6) at baseline. However, complicating the interpretation of the latter findings is the fact that those three scales correlate substantially with *SF-36* subscales measuring mental health. Are those subscales responding to mental health, which was very low at baseline for the depressed patients that Simon and his colleagues described?

Taft also questioned the validity of PCS and MCS as outcome measures on the basis of scores reported by Simon and his colleagues at the time of their 3-month follow-up. How well did the PCS summary measure agree with the PF subscale and how well did the MCS summary agree with the MH subscale at the 3-month follow-up in that study? Upon close inspection, it is clear that those depressed patients scored close to normal in terms of physical health at follow-up according to *both* the PCS summary and PF subscale (standardized NBS scores: PF = 51.8 and PCS = 50.7, with a norm of 50 for both). In our re-analysis of their published data, those patients also scored nearly normal at follow-up on the *oblique* SF-36 physical health factor score (i.e., 49.7 using norm-based scoring). At follow-up, their average mental health scores were well below the norm of 50 for all three mental health measures (i.e., MH = 47.3, MCS = 46.4 and 46.7 for the oblique mental health factor score). Thus, for both orthogonal and oblique scoring methods for the summary measures, these patients ended up at or above the norm for physical health and about 2-3 tenths of a standard deviation below the norm for mental health. In summary, the PF and MH subscales (the purest physical and mental health

subscales in the SF-36) showed the same pattern of results at follow-up as was shown by the PCS and MCS summary measures in the Simon et al study. (Our estimations of oblique summary scores, not reported by Simon et al, can be replicated by entering their published data into the SF-36 scoring utilities and by requesting "oblique" summary scores on the Internet at www.sf-36.com/nbs.

There is an important lesson from the example of "extreme" differences in physical and mental health reported by Simon et al. At baseline, average mental health scores were very low (more than two standard deviations below the population norm), even for depressed patients. The orthogonal MCS and the oblique mental health factor score agreed in this regard. Further, scores improved to just below normal (about one-quarter SD) or just above for mental and physical health, respectively. As Simon et al note, under these "extreme" circumstances the information value of changes in the profile of *SF-36* subscales may be greater and both subscales and summary measures should be thoroughly evaluated before conclusions are drawn. However, we disagree with Taft's conclusion that the PCS and MCS are, therefore, less valid summary measures.

What Do Studies of Treatment Outcomes Show?

We are aware of more than 120 publications about studies using the PCS and MCS However, we need not limit ourselves to these studies in our search for clues from the published literature. 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 can be considered. For purposes of evaluating the validity of PCS and MCS, we have focused our ongoing review on the more than 250 longitudinal studies reporting *SF-36* profiles and/or summary measures before and after treatment. For all of these studies,

we have standardized scores for the PCS and MCS summary measures and eight subscales using readily available *SF-36* Norm-Based Scoring (NBS) software. Using this scoring utility, *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 the U.S. and other countries on the Internet at www.sf-36.com/nbs.

Before commenting on the findings form these outcomes studies, we note that summary health measures should not agree completely with the subscales that they aggregate because they represent different "formulas" for specifying health. The same logic applies to comparisons between subscales in the *SF-36* profile. Pairs of subscales should not always agree because they measure different domains of health. For that reason, some "discrepancies" in results constitute support for the validities of different subscales. Again, that is why our manuals and articles documenting the summary measures recommend comparing results across subscales and summaries before drawing conclusions. Further, readers should be informed as to whether conclusion would have been different if based on subscale scores or summary measures.

In selecting a relatively small number of outcomes studies for discussion here, we looked for published examples of different patterns, including extreme differences in results for physical and mental health summary measures and/or subscale scores. We offer brief summaries and comment on results for the following four examples:

(1) Nearly equal treatment effects across the four most *physical SF-36* subscales (PF, RP, BP & GH) and the four most *mental* subscales (MH, RE, SF,& VT) health clusters of subscales in a cross-over placebo controlled study of medication for gastro-oesophageol reflux disease (Watson et al, 1997).

(2) Treatment effects predominantly in the four *SF-36* subscales in the *physical* health cluster (PF, RP, BP & GH); randomized trial comparing medications for rheumatoid arthritis (Kosinski et al, 2001, in press)

(3) Treatment effects predominantly in the four *SF-36* subscales in the *mental* health cluster for patients scoring extremely low across those subscales; randomized trial of medication for unipolar major depression (Heiligenstein et al, 1995; Beusterien et al, 1996), and

(4) Treatment effects predominantly in the *SF-36 physical* health cluster of subscales among patients scoring extremely low in those subscales and nearly normal on the MH subscale at baseline. Before and after comparison of consecutive patients who underwent either a hip or knee arthroplasty (Benroth and Gawande, 1999).

Each of these is discussed briefly below along with figures comparing subscale profiles and summary measures standardized using NBS.

First, we reiterate Taft's hypotheses: PCS and MCS scoring algorithms cause over- and under-estimation of physical and mental health status throughout the score range. Thus, treatments that improve only one component of health (physical or mental) will appear to negatively impact the other component if estimated using *SF-36* PCS or MCS summary measures. Likewise, Taft hypothesizes that declines in one component will cause over-estimation of improvements in the other component. Although Taft did

not speculate regarding what would happen if treatment improved both components of health, it follows from our understanding of Taft's hypothesis model that equivalent improvements across the eight *SF-36* subscales would cancel each other out when summarized using the PCS and MCS. We looked for evidence simply by reviewing the many outcomes studies from the literature and compared results for profiles and summary measures. We thank Taft for these testable hypotheses.

Inherent in comparisons between profiles and summary measures is the notion that the *SF-36* profile will give the "correct answer," although one that is a challenge to interpret because the profile gives eight answers and most are substantially inter-correlated. By comparing results for key subscales in the profile (e.g. PF and MH) with results based on PCS and MCS we can detect the hypothesized biases in the summary scores. We also recommend the reverse logic. Inspection of results for the summary measures is often very useful in drawing conclusions regarding the implications of the pattern of differences in the profile of subscales.

Roughly Equivalent Physical and Mental Health Improvements. Watson et al (1997) reported *SF-36* profile scores before and after four weeks of medication treatment for 12 patients with positive reflux symptom scores. As shown in Figure 1 below, average scores for these patients were below average for all subscales before treatment, more than 1 SD below average for RP, BP, GH and SF (which was the lowest subscale, nearly 2 SDs below average). NBS scores for PF and MH were very similar before treatment (44.9 and 43.3, respectively). As shown in the figure, PCS and MCS scores were 39.6 and 41.0, on average, at baseline, about one SD below their norms. Averages for both summaries were below the norm and below the averages for the PF and MH

subscales at baseline. The oblique summary scores averaged about 1-2 points lower. If negative coefficients for subscales aggregated in scoring either the PCS or MCS summary measure had the "canceling" effects hypothesized by Taft, we would expect higher average scores for PCS and MCS, in comparison with averages for subscale scores.

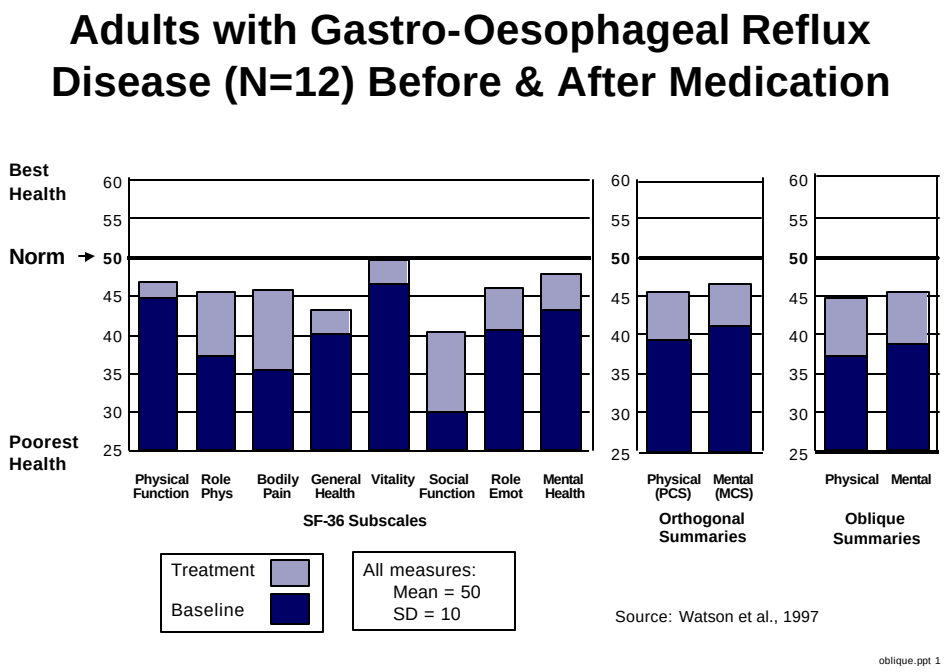

Figure 1

At follow-up, improvements were substantial in both sets of subscales (ES = 0.21 to 1.00 for subscales in the physical cluster and 0.31 to 0.69 for the mental cluster). Consistent with these improvements for the subscales, ES estimates for PCS and MCS were 0.59 and 0.54, respectively. ES estimates were even larger for the two oblique factor scores (0.71 & 0.68, respectively). We find from this example, and other studies with roughly equivalent physical and mental health improvements, no support for the "cancellation" effect hypothesized by Taft. Whether the lower oblique factor scores at

baseline and larger ES estimates at follow-up reflects greater validity or is the result of counting the same improvement twice requires further analyses, which are in progress.

Greater Physical than Mental Health Improvements. Kosinski et al (2001, in press) reported SF-36 profile and summary scores before and after 52 weeks of medication treatment for 424 patients with rheumatoid arthritis. Their results illustrate a study of improvements in both components of health but with extremely low baseline scores and average improvement in one component, in this case the physical component. The issues are whether estimates of mental health burden at baseline, based on MCS, are underestimated and whether mental health outcomes are cancelled or minimized, as hypothesized by Taft. As shown in Figure 2, average scores for these patients were below the norms for all subscales before treatment and more than 1 SD below the norm for six of the eight subscales (all but RE and MH). The PF score was extremely low at baseline (more than 2 SD's below the norm). The subscale scores for PF and MH were very different before treatment (28.6 and 44.9, respectively). At baseline, PCS and MCS summary scores were very consistent with this patter (28.4 and 46.8, respectively). If negative coefficients for the PCS summary measure had the "canceling" effect hypothesized by Taft, we would expect a substantially higher average score for MCS, in comparison with the MH subscale. To the contrary, as shown in the figure, average MCS and MH scores were 46.8 and 44.9, respectively, at baseline. The oblique summary physical health score was nearly identical to PCS at baseline (28.39 and 28.42, respectively). However, the oblique summary mental health score at baseline was more than one-half SD below both the MCS and MH subscale.

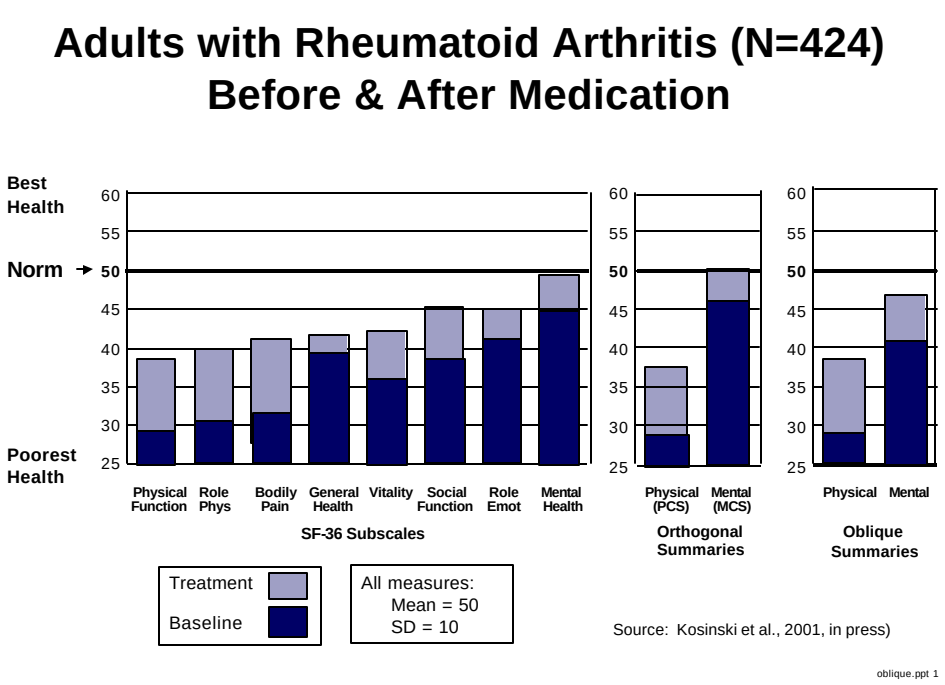

Figure 2

Average improvements with treatment were substantially larger for three of the physical health subscales (PF, RP and BP) (ES = 0.85 to 0.91) in comparison with two of the mental health scales (MH and RE) (ES = 0.46 and 0.38, respectively). The effect of treatment on PCS was much larger than on MCS (ES = 0.90 versus 0.35, respectively) and the improvement in the oblique summary physical health score agreed with the MCS (ES = 0.96 and 0.90, respectively). The oblique summary mental health score improved more than MCS (ES = 0.57 versus 0.35), an effect size noticeably more than that observed for the MH and RE subscales (0.46 and 0.38, respectively). We find from this study no support for the "cancellation" effect hypothesized by Taft. The estimate of the mental health burden of RA at baseline, as estimated with MCS, was not minimized by the substantial physical health burden and the substantially larger improvement in physical health estimated with PCS did not cancel out the improvement in mental health

estimated with MCS. This study suggests that when scores are very low for one component of health, summary measures based on oblique factors underestimate scores for the other component. This finding warrants further study.

Severe Physical Impairment and Near-Normal Mental Health. Benroth and Gawande (1999) reported *SF-36* profile scores before total hip or total knee arthroplasty and again one year postoperatively. Published *SF-36* (0-100) scores reported in their article were standardized and PCS and MCS summary scores (not reported by the authors) were also estimated using the *SF-36* NBS software. This study illustrates an interpretation of disease burden and treatment outcomes complicated by an extreme preoperative decrement in physical health, near-normal mental health (MH subscale) and near-normal self-evaluated health (GH subscale), both within 2.5 points of their norms. In addition, the other three subscales in the physical health cluster (PF, RP & BP) all show substantial decrements in physical health (-2 to -2.4 SDs below norms). Are the decrements in the VT, SF and RE subscales indicative of poor mental health (in contrast to the near-normal MH subscale score) or do they indicate poor physical health, or both? The substantial correlations among the subscales add to the challenge of interpreting this pattern of results.

If negative coefficients for subscales in the physical health cluster had the "canceling" effect hypothesized by Taft, we would expect an inflated MCS score, in comparison with the four subscales in the mental health cluster. To the contrary, as shown in Figure 3, the average MCS score was very close to the norm (48.5) and in rough agreement with the MH subscale (47.6). The average PCS score at baseline was

extremely low (29.1, -2.1 SD), as expected and nearly identical to the estimate based on the oblique factor score (-2 SD).

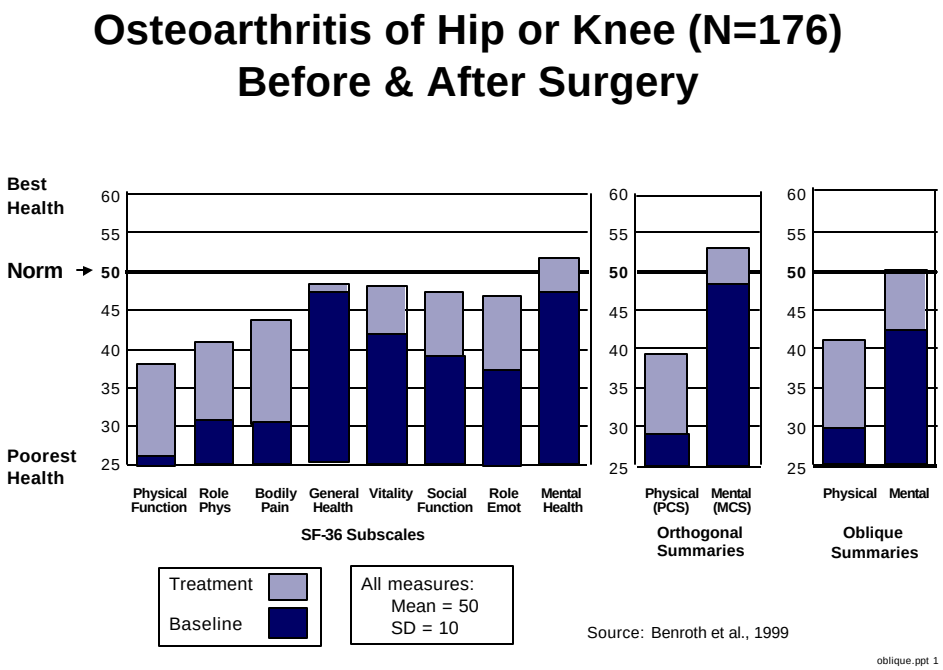

Figure 3

One year after surgery, all subscales in the mental health cluster improved to within about 2-3 points of the norm (below or above) and the MCS improved to three points above the norm. PCS improved substantially, on average (ES = 1.04) to within one SD of the norm, as did the oblique physical factor score (ES = 1.12). In contrast to what would be expected due to bias introduced by negative coefficients, the ES estimates for the MCS and the oblique mental factor score were nearly identical 0.67 and 0.65 SD units, respectively) and agreed with that for the MH subscale (ES = 0.69). The most obvious discrepancy in the results from this study is that between the baseline score estimates for the two mental summaries; the oblique mental factor score estimate was much lower than the MCS and MH subscale at baseline (42.9 versus 48.5 and 47.6,

respectively). We find no support for the "cancellation" effect hypothesized by Taft in the results from this study. This study may be another example of a downward bias in the oblique mental factor score estimate, in relation to MCS and the MH subscale, caused by the substantial correlation between the oblique physical and mental health factors. Again, this hypothesis warrants further study.

Severe Mental Health Decrements. Heiligenstein et al (1995) reported *SF-36* profile scores before and after six weeks of medication treatment for 261 elderly patients with unipolar depression, in comparison with randomized controls. A companion article (Beusterien et al, 1997) compared results for both SF-36 subscales and summary measures with those based on parallel clinical measures of depression, including the Geriatric Depression Scale (GDS), the Hamilton Depression Rating Scale (HAM-D), and the Clinician's Global Impression (CGI) of depression severity. This is an example of a study in which baseline decrements in the *SF-36* profile and treatment effects were extreme for the cluster of mental healthy subscales.

As shown in Figure 4, average scores for these depressed patients were below the norms for all subscales before treatment, and particularly so (more than 1 SD) for the four mental health subscales (MH, RE, SF & VT). The NBS score for PF was much higher (about 1 SD) in comparison with MH at baseline (42.2 and 32.8, respectively). At baseline, PCS and MCS summary scores showed a very similar pattern with PCS much higher than MCS (46.5 and 31.5, respectively). The oblique summary mental health score was nearly identical to the MCS at baseline (31.2 and 31.5, respectively). The oblique physical factor score was one-half SD below PCS at baseline (41.5 versus 46.5, respectively).

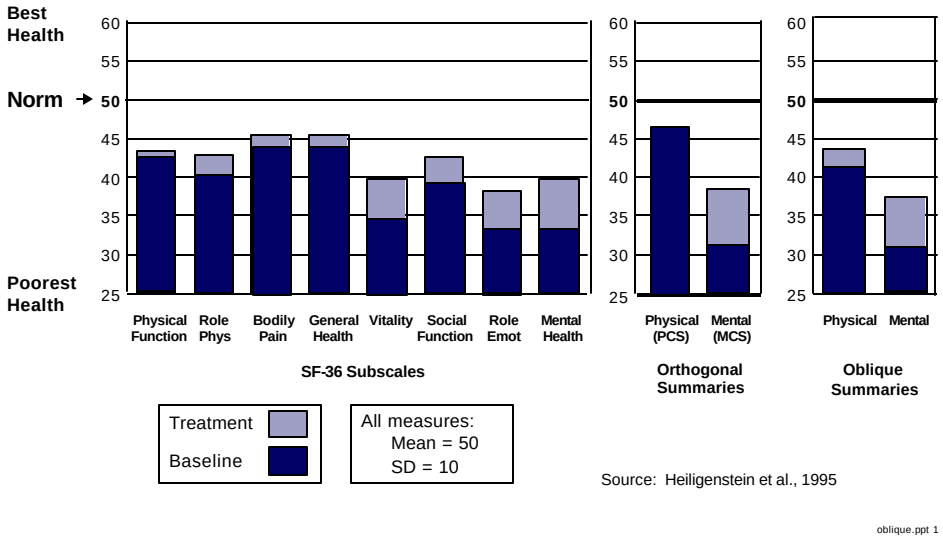

Figure 4

Average improvements with treatment were larger for the four mental health subscales (ES = 0.31 to 0.69) in comparison with the four physical health subscales (0.12 to 0.30). The effect of treatment on MCS was very much larger than the effect on PCS (0.67 versus 0.02) and the oblique mental health factor score agreed with the MCS (ES = 0.65 and 0.67, respectively). The oblique physical health factor score improved much more than the PCS (0.20 versus 0.02, respectively) and noticeably more than the ES units observed for three of the four physical health subscales (0,12 to 0.19), the exception being BP (ES = 0.30). Clearly, PCS estimates of change were not biased by changes in mental health in this trial. MCS improved substantially in this trial, as it has in numerous other published trials showing effects on the mental health cluster of *SF-36* subscales. Further, this trial illustrates an advantage of the MCS in summarizing mental health outcomes. MCS was substantially (about 35%) more valid than the best subscale (F =

100.0 and F = 73.96 for MCS and MH, respectively) in longitudinal tests of validity in discriminating between responders and non-responders (defined using the HAM-D) and about 40 % more valid (F= 49.9 versus 35.6, respectively) in the test of validity in discriminating across levels of severity (defined using the CGI). A similar advantage of PCS over subscale scores in detecting treatment improvements has been noted in a clinical trial of adults with asthma (Ware, Kemp, Buchner, et al., 1998).

We find from this example, and other studies with similar outcomes, strong support for the discriminant validity and responsiveness of the MCS in tests based on accepted clinical standards. This study also illustrates that when effects are consistent across the subscales in the mental health cluster, the MCS summary performs substantially better than any one of those subscales thereby reducing the number of comparisons and increasing responsiveness. We find no support for Taft's hypotheses in the results from this study.

Final Comments as Published in *QOLR*

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 www.sf-36.com). 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. 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 Norway, 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.