J Clin Epidemiol Vol. 51, No. 11, pp. 1171–1178, 1998

ELSEVIER

0895-4356/98 $–see front matter
PII S0895-4356(98)00109-7

Cross-Validation of Item Selection and Scoring for the SF-12 Health Survey in Nine Countries: Results from the IQOLA Project

Barbara Gandek,[1,*] *John E. Ware,*[1] *Neil K. Aaronson,*[2] *Giovanni Apolone,*[3] *Jakob B. Bjorner,*[4] *John E. Brazier,*[5] *Monika Bullinger,*[6] *Stein Kaasa,*[7] *Alain Leplege,*[8] *Luis Prieto,*[9] *and Marianne Sullivan*[10]

[1]Health Assessment Lab at the Health Institute, New England Medical Center, Boston, Massachusetts; [2]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands; [3]Dipartimento di Oncologia, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy; [4]Institute of Public Health, University of Copenhagen, Copenhagen, Denmark; [5]Sheffield Health Economics Group, School of Health and Related Research, University of Sheffield, Sheffield, United Kingdom; [6]Abteilung Für Medizinische Psychologie, Universitätskrankenhaus Eppendorf, Hamburg, Germany; [7]Unit for Applied Clinical Research, The Norwegian University for Science and Technology, Trondheim, Norway; [8]Institut National de la Santé et de la Recherche Médicale (INSERM) Unité 292, Hôpital de Bicêtre, Le Kremlin-Bicêtre, France; [9]Health Services Research Unit, Institut Municipal d'Investigació Mèdica (IMIM), Barcelona, Spain; and [10]The Health Care Research Unit, Institute of Internal Medicine, Sahlgrenska University Hospital and Göteborg University, Göteborg, Sweden

ABSTRACT. Data from general population surveys (n = 1483 to 9151) in nine European countries (Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom) were analyzed to cross-validate the selection of questionnaire items for the SF-12 Health Survey and scoring algorithms for 12-item physical and mental component summary measures. In each country, multiple regression methods were used to select 12 SF-36 items that best reproduced the physical and mental health summary scores for the SF-36 Health Survey. Summary scores then were estimated with 12 items in three ways: using standard (U.S.-derived) SF-12 items and scoring algorithms; standard items and country-specific scoring; and country-specific sets of 12 items and scoring. Replication of the 36-item summary measures by the 12-item summary measures was then evaluated through comparison of mean scores and the strength of product-moment correlations.

Product-moment correlations between SF-36 summary measures and SF-12 summary measures (standard and country-specific) were very high, ranging from 0.94–0.96 and 0.94–0.97 for the physical and mental summary measures, respectively. Mean 36-item summary measures and comparable 12-item summary measures were within 0.0 to 1.5 points (median = 0.5 points) in each country and were comparable across age groups.

Because of the high degree of correspondence between summary physical and mental health measures estimated using the SF-12 and SF-36, it appears that the SF-12 will prove to be a practical alternative to the SF-36 in these countries, for purposes of large group comparisons in which the focus is on overall physical and mental health outcomes. J CLIN EPIDEMIOL 51;11:1171–1178, 1998. © 1998 Elsevier Science Inc.

KEY WORDS. Construct validity, health status indicators, SF-36 Health Survey, translations, cross-cultural comparisons, SF-12 Health Survey

INTRODUCTION

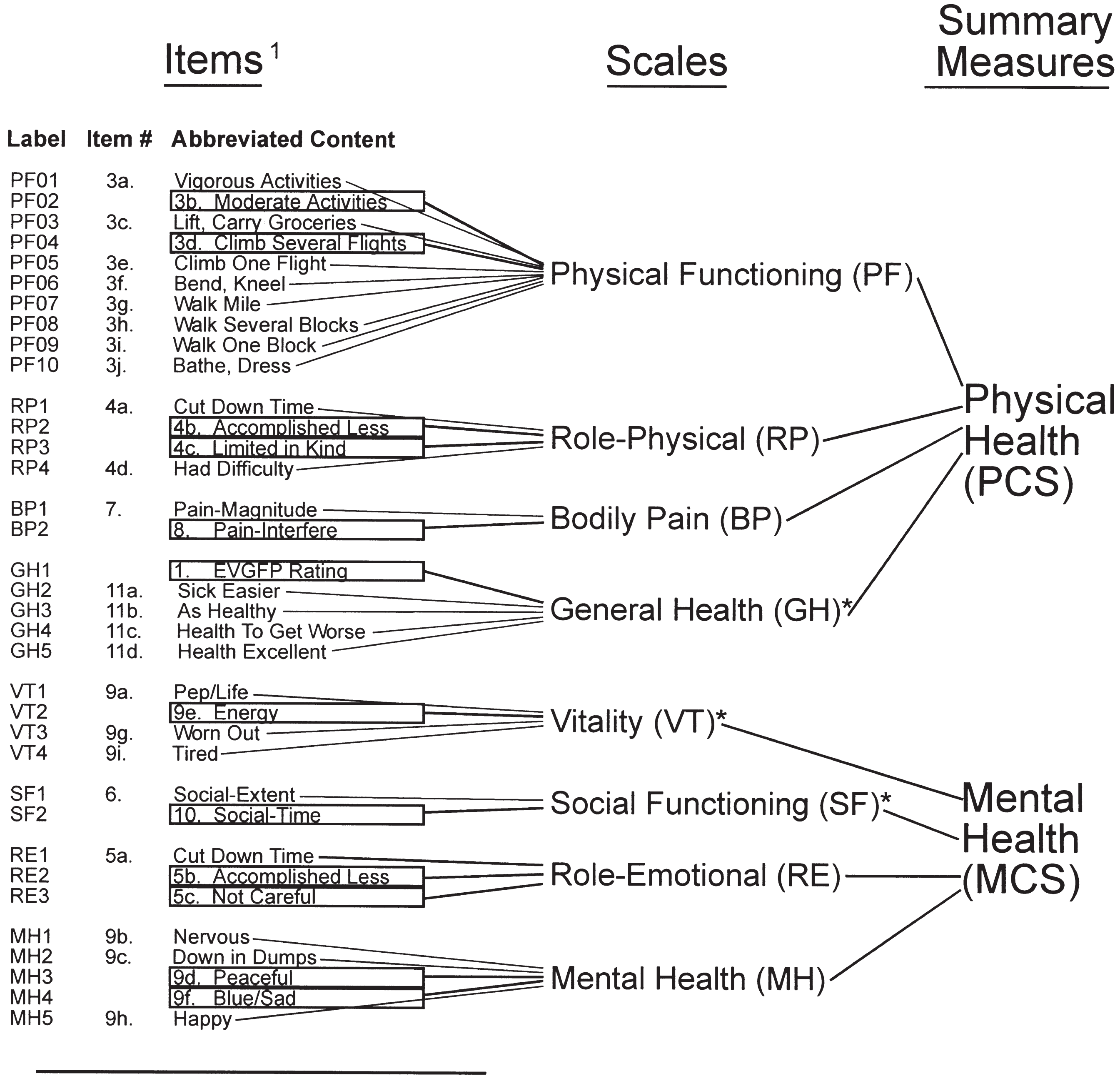
Although the 36-item SF-36 Health Survey is a short-form measure, for some applications even a questionnaire with 36 questions is too lengthy. Large general population surveys may only have room for one page of questions about health. Questionnaires that include disease-specific measures may not have room for a generic measure of health status such as the SF-36. In addition, although the SF-36 can be completed in a relatively short amount of time (5 to 10 minutes on average), this may be too great a burden for some respondents. Therefore, use of a shorter form than the SF-36 is warranted in a number of instances.

Development of two summary measures from the SF-36 [1–5] suggested that it might be possible to develop a shorter survey which would reproduce the SF-36 physical and mental health summary measures with fewer items. Because the number of items in a survey is dependent on the

[*]Address for correspondence: Barbara Gandek, M.S., Health Assessment Lab, 750 Washington Street, NEMC #345, Boston, MA 02111.

Accepted for publication on 7 July 1998.

number of dimensions for which scores are to be estimated, fewer questions are needed to calculate two summary scores than to calculate eight scale scores. Thus, the SF-12 Health Survey was originally developed in the United States to provide a shorter alternative to the SF-36, for use in large-scale health measurement and monitoring efforts in which a 36-item questionnaire was too lengthy and in which the focus was on overall physical and mental health outcomes [6,7]. The SF-12 contains a subset of 12 items from the SF-36, including one or two items from each of the eight SF-36 scales (Figure 1). Two items are included from the Physical Functioning and Mental Health scales because



* Significant correlation with other summary measure.

[1] Items in boxes were selected for SF-12.

Adapted from [7]

FIGURE 1. SF-12 measurement model.

these scales have been shown to best predict physical and mental health; two items each are also included from both Role Functioning scales, because these are relatively coarse scales. One item each is included from the remaining four scales. Information from all 12 items is used to construct physical and mental component summary measures (PCS-12 and MCS-12).

In the U.S. general population, the SF-12 items explained more than 90% of the variance in the SF-36 physical (PCS-36) and mental (MCS-36) summary measures [6]. In cross-validation with data from the Medical Outcomes Study, the PCS-36 and PCS-12 correlated 0.95 and the MCS-36 and MCS-12 correlated 0.97. Within the U.S. general population, mean PCS-36 and PCS-12 scores were within 1 point across subgroups differing in age and gender, and similar results were found in comparing the MCS-36 and MCS-12 [7]. Expected relationships between the SF-12 summary measures and clinical criteria were verified. Thus, in the United States the SF-12 reproduced the SF-36 summary measures with the same interpretations.

The two-component model of physical and mental health has been replicated using SF-36 data from large general population samples in nine Western European countries studied to date (Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom) [8]. These findings supported the derivation and testing of SF-36-based physical and mental health summary measures in these countries [9]. In this study, we cross-validate the selection of questionnaire items for the SF-12 in these nine countries and examine how well the SF-12-based summary measures reproduce the SF-36-based summary measures. We also compare the use of country-specific versus standard (U.S.-derived) scoring algorithms for the SF-12 summary measures.

METHODS

Data

Data come from 10 general population surveys, which have been described in detail elsewhere [10]. In brief, samples were selected to be nationally representative in nine countries (Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, the United Kingdom, and the United States). Data from Sweden were collected through seven mail surveys conducted in various regions of Sweden to provide a broad cross-section of the population [11]. Self-administration of the SF-36 was used in all countries, with the exceptions of Italy (50% personal interview), Spain (100% personal interview), the United Kingdom (100% personal interview), and the United States (32% telephone interview). Twenty-seven percent of Italian respondents received an explanation of one or more questions from the interviewer (23%) or other person (4%); these respondents were not included in the analysis, due to concerns about possible bias [12]. Response rates ranged from 61–81%. The mean age ranged from 41.1 years to 47.6 years; slightly more than half of the respondents were female in each country except in the Netherlands in which the sample was 44% female. Data completeness was satisfactory, and Cronbach's alpha for the eight SF-36 scales ranged from 0.68 to 0.94, with a median value of 0.83 [3].

Health Status Measures

Summary physical (PCS) and mental (MCS) component scores were constructed from the SF-36 and SF-12, using standard (U.S.) and country-specific scoring algorithms. Five sets of PCS/MCS summary components were derived in each country, as noted in Table 1. Two sets of SF-36 summary component scores were calculated. One set used

TABLE 1. Description of SF-36 and SF-12 physical (PCS) and mental (MCS) summary measures

Label	Calculated from	Weights used	Normed to
PCS-36/MCS-36	Eight SF-36 scales	U.S. factor weights	Mean = 50 and SD = 10 in U.S. general population
CPCS-36/CMCS-36	Eight SF-36 scales	Country-specific factor weights	Mean = 50 and SD = 10 in country sample
PCS-12/MCS-12	12 items from the standard U.S. SF-12	U.S. regression weights	Mean = 50 and SD = 10 in U.S. general population
CPCS-12/CMCS-12	12 items from the standard U.S. SF-12	Country-specific regression weights	Mean = 50 and SD = 10 in country sample
CSPCS-12/CSMCS-12	12 items selected separately for each country	Country-specific regression weights	Mean = 50 and SD = 10 in country sample

standard U.S. scoring algorithms and was normed to the U.S. general population (PCS-36/MCS-36) [13]. The second set used previously derived country-specific scoring algorithms and was normed to country-specific general population samples (CPCS-36/CMCS-36)[9].

Three sets of SF-12 physical and mental summary measures were calculated. All three methods applied regression weights to the 12 items, using separate physical and mental regression weights for each item response category. The first set of summary measures (PCS-12/MCS-12) was calculated using the standard SF-12 items and U.S. regression weights [7]. Scores were transformed to have a mean of 50 and standard deviation of 10 in the U.S. general population.

The second set of SF-12 summary scores (CPCS-12/CMCS-12) used the standard U.S. SF-12 items and country-specific regression weights. These weights were derived from regressions that used the response categories for the 12 items as independent variables and CPCS-36 and CMCS-36 as the dependent variables, to arrive at a set of physical regression weights and mental regression weights in each country. Scores were transformed to have a mean of 50 and a standard deviation of 10 in each country.

The third set of SF-12 summary scores (CSPCS-12/CSMCS-12) used the 12 items in each country that were the best predictors of the country-specific PCS-36 and MCS-36, following methods used to construct the SF-12 in the United States [6,7]. After first confirming that the two-component factor structure of the SF-36 was replicated in each country, forward stepwise regression was used, with CPCS-36 and CMCS-36 as the dependent variables, to identify a subset of SF-36 items which explained at least 90% of the variance in the CPCS-36 and in the CMCS-36. SF-36 items were not recoded prior to entry into the regression. The best predictors of the CPCS-36 and the CMCS-36 were combined to derive the country-specific SF-12 items for each country. Item selection was made using the constraints that at least one item must be selected for each of the eight SF-36 concepts, and that two items would be selected from the Physical Functioning, Mental Health, Role-Physical, and Role-Emotional scales, for the same conceptual reasons as in the derivation of the U.S. SF-12. In addition, the overall health item ("in general, would you say your health was: excellent to poor") was used in all country-specific questionnaires, due to its widespread use as a single-item measure in numerous health surveys. Because the five general health items had partial R^2 values ranging from 0.00 to 0.03 across the PCS and MCS regressions in all countries, selection of the overall health item instead of another general health item did not affect the percentage variance explained by the selected 12 items in any country. After selection of the 12 items within each country, country-specific regression weights were derived and were used to calculate the SF-12 summary scores. The scores were transformed to have a mean of 50 and a standard deviation of 10 in each country.

Analyses

We evaluated how well the SF-12 replicated the SF-36 summary measures by examining the proportion of variance, or R^2, in PCS-36 and MCS-36 scores that was explained by the 12 items; we hypothesized that this would be 90% or greater. In addition, we examined the correlations between SF-36 and SF-12 summary measures (e.g., PCS-36 with PCS-12, CPCS-36 with CPCS-12, CSPCS-36 with CSPCS-12), which also were expected to be high. We also examined the correlations between pairs of physical and mental summary measures that were scored using the same method (e.g., PCS-12 and MCS-12); we hypothesized that these correlations would be positive and low.

We compared descriptive statistics (means and standard deviations) for the SF-36 and SF-12 summary measures, scored using standard (U.S.) scoring algorithms, to determine how closely the SF-12 measures replicated the SF-36 measures on average, both overall (for respondents age 18–74) and by age group (18–44, 45–64, 65–74). Data from the United States was included in these comparisons. No adjustment was made for other differences among countries (e.g., gender, presence of chronic conditions). Based on previous studies in the United States [3, 6, 7], we hypothesized that mean PCS-36 and PCS-12 scores would be similar in each country and would decline with age; that mean MCS-36 and MCS-12 scores would be similar in each country and would stay stable or increase slightly with age; and that the amount of variance in PCS scores explained by age would be slightly lower for the SF-12 measures than for the SF-36 summary measures, due to reduced precision of the 12-item measures.

Because one goal of the analysis was to select the 12 items that best predicted the SF-36 summary measures in each country, item selection was made using a subset of data from each country. If the 12 items selected within each country differed greatly from the standard set of 12 items, or if results using standard versus country-specific scoring differed greatly, further evaluation could be done on the remaining data. Therefore, development of country-specific SF-12 scoring algorithms and correlational analyses were based on a random two-thirds sample of each dataset. Comparisons of mean scores are based on the entire sample in each country, but are limited to those respondents for whom both 36-item and 12-item summary measures could be calculated.

RESULTS

The 12 items selected in each European country to empirically reproduce the SF-36 summary measures agreed considerably with the standard SF-12 items selected in the United States (Table 2; verbatim item content is provided elsewhere in this issue [14]). In 91 of 108 instances across the nine countries, the country-specific items were the same as

TABLE 2. Twelve items that best reproduce the SF-36 PCS/MCS measures in ten countries

Scale	U.S.	Denmark	France	Germany	Italy	Netherlands	Norway	Spain	Sweden	U.K.
PF	PF02	•	•	•	•	•	•	PF08	•	•
	PF04	•	•	PF07	•	•	•	•	•	PF07
RP	RP2	•	•	•	•	•	•	•	•	•
	RP3	•	•	•	•	•	RP4	RP4	•	•
BP	BP2	•	•	•	•	•	•	•	•	•
GH	GH1	•	•	•	•	•	•	•	•	•
VT	VT2	VT4	VT4	•	VT4	VT4	•	VT1	VT4	VT3
SF	SF2	SF1	•	•	SF1	•	SF1	•	SF1	SF1
RE	RE2	•	•	•	•	•	•	•	•	•
	RE3	•	•	•	•	•	•	•	•	•
MH	MH3	•	•	•	•	•	•	•	•	•
	MH4	•	•	•	•	•	•	•	•	•

•Indicates same item as in U.S. derived SF-12.

Abbreviations: PF = Physical Functioning; RP = Role-Physical; BP = Bodily Pain; GH = General Health; VT = Vitality; SF = Social Functioning; RE = Role-Emotional; MH = Mental Health

in the standard U.S. SF-12. The same two Physical Functioning items, PF02 (moderate activities) and PF04 (several flights of stairs), were the two best predictors of the country-specific CPCS-36 in six countries, and one of these two items was included in the other three countries. The Bodily Pain item that measures limitations in normal work due to pain (BP2) entered the PCS stepwise regression as the first or second variable in every country. The same Role-Physical and Role-Emotional items generally were the best predictors of CPCS-36 and CMCS-36, respectively. The two best Mental Health items in the United States, MH3 (calm and peaceful) and MH4 (downhearted and blue), were the best predictors of the CMCS-36 in every country. While the Vitality item selected for the U.S. SF-12 was "full of energy," Vitality items measuring fatigue (VT3 or VT4) were selected in all other countries except Germany and Norway. The greatest number of differences from the U.S. standard solution was seen in Spain and the United Kingdom. The 12 items selected within each country explained 88–92% of the variance in CPCS-36 scores and 89–94% of the variance in CMCS-36 scores.

The standard SF-12 items explained 89–92% of the variance in PCS-36 scores and 88–94% of the variance in MCS-36 scores, across all nine countries. Correlations between the SF-12 summary measures scored using standard items and weights, and the SF-36 summary measures scored using standard scoring algorithms (PCS-12 with PCS-36; MCS-12 with MCS-36) were very high in all countries, ranging from 0.94–0.97 (Table 3). Correlations between the SF-12 summary measures scored with standard items and country-specific weights (CPCS-12 and CMCS-12) and the SF-36 summary measures scored using country-specific scoring algorithms (CPCS-36 and CMCS-36) ranged from 0.94–0.97 across countries. Similarly, correlations between the SF-12 summary measures scored with country-specific items and country-specific weights (CSPCS-12 and CSMCS-12) and the SF-36 summary measures scored using country-specific scoring algorithms (CPCS-36 and CMCS-36) also were high, ranging from 0.94–0.97. Within each country, the correlations among the physical summary measures were within 0.01 of each other, as was the case for the mental summary measures.

Correlations between pairs of SF-12 physical and mental summary measures scored using standard (PCS-12 with MCS-12) and country-specific (CPCS-12 with CMCS-12) algorithms generally were positive and low (Table 4). The

TABLE 3. Correlations between SF-36 and SF-12 summary measures in nine countries

Country	*n*	PCS-36/ PCS-12	CPCS-36/ CPCS-12	CPCS-36/ CSPCS-12	MCS-36/ MCS-12	CMCS-36/ CMCS-12	CMCS-36/ CSMCS-12
Denmark	2746	0.95	0.94	0.94	0.96	0.95	0.95
France	2455	0.94	0.94	0.94	0.96	0.96	0.96
Germany	1955	0.96	0.96	0.96	0.94	0.95	0.95
Italy	985	0.94	0.94	0.94	0.95	0.94	0.94
Netherlands	1171	0.95	0.95	0.95	0.97	0.97	0.97
Norway	1515	0.95	0.96	0.96	0.97	0.97	0.97
Spain	6124	0.94	0.94	0.94	0.94	0.94	0.94
Sweden	5970	0.95	0.94	0.94	0.97	0.97	0.97
United Kingdom	1372	0.96	0.95	0.96	0.94	0.94	0.95

TABLE 4. Correlations between SF-36 and SF-12 physical and mental summary measures using standard and country-specific scoring

Country	PCS-12/MCS-12	CPCS-12/CMCS-12
Denmark	0.06	0.04
France	0.08	0.06
Germany	0.06	0.02
Italy	0.20	0.10
Netherlands	0.13	0.05
Norway	0.08	−0.02
Spain	0.17	0.06
Sweden	0.15	0.12
United Kingdom	0.19	0.02

correlations between SF-12 physical and mental summary measures ranged from −0.02–0.20 (median = 0.07) and all but one were positive.

Within each country, unadjusted mean scores for the PCS-36 and PCS-12, and also for the MCS-36 and MCS-12, generally were within 1 point (Table 5). (All scores were calculated using standard U.S. scoring algorithms, which yield a mean score of 50 and standard deviation of 10 in the U.S. general population.) Analysis of difference scores (e.g., PCS-36 minus PCS-12) also showed similar results (data not reported). Within each age group in each country, mean PCS-36 and PCS-12 scores were within 1 point of each other in 23 of 30 comparisons, and were within 1.7 points in all comparisons (Table 6). Mean MCS-36 and MCS-12 scores were within 1 point of each other in 25 of 30 comparisons and were within 1.4 points in all comparisons. As hypothesized, physical summary scores declined with age in all countries. Mental summary scores generally remained stable or increased slightly with age, although mental health scores declined with age in Italy and Spain. F-statistics, which measure the amount of separation in scores between age groups relative to the within group (error) variance, generally were slightly lower for the PCS-12 than the PCS-36 in each country, with some exceptions.

DISCUSSION

In each of the nine European countries, there were substantial correlations between the summary measures scored from the SF-36 and SF-12 Health Surveys. Correlations were also substantial between scores based on three different estimation methods (standard items and scoring weights; standard items and country-specific scoring weights; and country-specific items and scoring weights). Mean scores were also very comparable across estimation methods. In addition, there was a high degree of replication in the selection of 12 items for the SF-12 across nine European countries and in comparison with items selected for the SF-12 in the United States. Thus, the SF-12 appears to provide good reproductions of the SF-36 summary measures in these countries.

TABLE 5. Mean PCS and MCS scores (standard deviation) in ten countries

Country	*n*	PCS-36	PCS-12	MCS-36	MCS-12
Denmark	3242	51.5	51.0	54.0	52.8
		(8.6)	(8.1)	(8.3)	(8.3)
France	2743	52.2	51.2	48.4	48.4
		(8.0)	(7.4)	(9.5)	(9.4)
Germany	2453	50.7	49.6	51.4	52.3
		(9.8)	(8.7)	(8.1)	(8.0)
Italy	1413	52.7	51.2	47.6	47.8
		(7.8)	(7.4)	(10.1)	(10.1)
Netherlands	1479	49.7	49.4	52.1	51.6
		(9.3)	(8.8)	(9.7)	(9.2)
Norway	1885	51.2	50.3	51.2	50.6
		(9.3)	(8.8)	(9.8)	(9.9)
Spain	8494	51.0	49.9	51.9	51.8
		(9.8)	(9.0)	(9.4)	(9.0)
Sweden	7175	50.8	50.3	53.5	52.9
		(9.1)	(8.5)	(10.0)	(9.6)
United Kingdom	1751	50.8	50.9	52.2	52.1
		(10.2)	(9.4)	(9.4)	(8.7)
United States	2105	50.8	50.8	50.0	50.0
		(9.4)	(8.9)	(9.9)	(9.5)

Note: Limited to adults age 18–74. All summary measures are calculated using standard (U.S.) scoring algorithms.

One limitation of this study is that it assumes that responses to SF-12 items that were interspersed within the SF-36 will be the same when those 12 items are administered alone. In support of this assumption, research in Australia found no significant differences in mean SF-12 summary scores when the SF-12 was embedded in the SF-36 and when the SF-12 was administered by itself to an equivalent and independent sample [15]. Similar results have been reported for the United States [7]. Thus, it seems reasonable to assume that similar conclusions will be reached in these nine European countries. This study also did not address the issue of the empirical validity of the SF-12. Studies of the validity of the SF-12 among groups known to differ in clinical profiles have been reported in the United States [6] and in the United Kingdom [16].

A question that often arises is whether standard (U.S.-derived [7]) or country-specific scoring algorithms are most appropriate for scoring the SF-12 in countries other than the United States. Based on the results reported here, we conclude that there is little difference and we recommend standard (U.S.-derived) scoring of the SF-12 summary measures, so that data can be compared and interpreted across countries in relation to standard benchmarks, namely scores with a mean of 50 and standard deviation of 10 in the U.S. general population. For comparisons within a country, standard and country-specific scoring are expected to lead to the same conclusions because of their high intercorrelations.

TABLE 6. Mean PCS and MCS scores (standard deviation) by age group in ten countries

Country	Age	*n*	PCS-36	PCS-12	MCS-36	MCS-12
Denmark	18–44	1945	53.6 (6.4)	53.0 (6.0)	53.6 (7.9)	52.3 (8.0)
	45–64	996	49.4 (9.8)	48.8 (9.4)	54.5 (8.8)	53.4 (8.7)
	65–74	301	45.2 (10.9)	44.9 (10.4)	55.5 (8.8)	54.1 (8.8)
	F		195.0	199.3	9.6	8.9
France	18–44	1508	54.3 (6.6)	52.9 (6.0)	48.1 (9.6)	48.4 (9.5)
	45–64	763	50.1 (8.7)	49.4 (8.0)	48.8 (9.4)	48.6 (9.4)
	65–74	472	46.1 (9.4)	45.7 (9.0)	48.9 (9.2)	48.3 (9.2)
	F		182.9	162.9	1.7	0.1
Germany	18–44	1209	54.1 (7.2)	52.5 (6.3)	50.9 (7.9)	52.1 (7.8)
	45–64	876	48.4 (10.4)	47.7 (9.5)	51.5 (8.3)	52.2 (8.2)
	65–74	368	43.4 (11.1)	43.5 (10.1)	53.4 (8.4)	53.4 (8.1)
	F		227.2	203.2	12.5	3.8
Italy	18–44	815	54.2 (6.5)	52.7 (6.0)	47.8 (9.9)	48.2 (9.8)
	45–64	479	51.0 (8.1)	49.7 (7.9)	47.1 (10.3)	47.5 (10.3)
	65–74	119	45.8 (10.9)	44.1 (10.7)	46.8 (11.7)	46.4 (11.6)
	F		73.9	61.9	0.9	1.8
Netherlands	18–44	764	52.5 (7.1)	51.7 (6.8)	51.6 (9.5)	51.4 (9.1)
	45–64	507	47.8 (10.0)	47.9 (9.6)	52.0 (10.1)	51.4 (9.6)
	65–74	208	44.5 (11.1)	45.2 (10.5)	53.6 (9.3)	52.9 (8.6)
	F		86.7	61.9	3.4	2.3
Norway	18–44	1144	53.3 (7.8)	52.2 (7.4)	50.3 (9.9)	49.9 (10.1)
	45–64	565	49.2 (10.0)	48.6 (9.3)	52.2 (9.4)	51.4 (9.5)
	65–74	176	43.9 (11.1)	43.3 (10.6)	54.0 (9.2)	53.1 (8.8)
	F		104.4	102.6	15.5	10.8
Spain	18–44	6548	54.3 (6.9)	52.8 (6.2)	52.3 (8.6)	52.5 (8.1)
	45–64	1253	48.5 (10.4)	47.8 (9.8)	51.7 (10.0)	51.5 (9.6)
	65–74	693	42.6 (11.8)	42.3 (11.1)	50.3 (11.0)	49.8 (10.7)
	F		987.8	909.1	22.4	44.5
Sweden	18–44	4386	52.7 (7.8)	52.0 (7.2)	53.1 (9.8)	52.6 (9.5)
	45–64	2116	48.8 (9.7)	48.6 (9.2)	54.1 (10.0)	53.3 (9.6)
	65–74	673	44.2 (10.9)	44.5 (10.4)	54.6 (10.7)	53.6 (10.3)
	F		353.3	309.0	11.0	5.9
United Kingdom	18–44	888	53.7 (7.5)	53.4 (7.0)	51.9 (8.6)	52.2 (7.7)
	45–64	588	48.7 (11.5)	49.1 (10.6)	51.8 (10.3)	51.4 (9.8)
	65–74	275	44.8 (12.1)	45.3 (11.2)	54.4 (9.4)	53.2 (9.1)
	F		100.3	91.2	7.5	4.2
United States	18–44	1123	53.1 (7.5)	52.9 (6.9)	49.2 (9.9)	49.5 (9.4)
	45–64	574	47.9 (10.5)	48.2 (10.2)	50.9 (9.8)	50.5 (9.7)
	65–74	408	43.5 (11.2)	43.7 (11.0)	52.6 (9.3)	52.1 (9.5)
	F		150.1	146.3	14.0	7.9

Note: All summary measures are calculated using standard (U.S.) scoring algorithms.

As demonstrated in this article, mean summary scores for adults age 18 to 74 in many European countries differed from the U.S. scores for both summary measures. Thus, these country differences in health are noteworthy and warrant further study. For this purpose, we recommend standard scoring coefficients, means and standard deviations, so that differences between countries are not transformed away. If country-specific scoring is used within a study conducted in one country, publications should precisely document the scoring that was used. Documentation of country-specific SF-12 scoring algorithms is forthcoming from the IQOLA Project (for more information, see www.iqola.org).

Another question that often arises is when to use the SF-12 rather than the SF-36. The SF-12 represents a calculated compromise between the objectives of practicality and the statistical precision of scores. The content validity of the SF-12 was enhanced by including one or two items from each of the eight health concepts in the SF-36. These 12 items also represent a variety of operational definitions of health, including what respondents are able to do, the distress and well-being they feel, how their everyday lives are affected, and how they evaluate their health status [7]. The SF-12 also meets the practical need for a health survey that can be printed on one page and can be administered in 2 minutes or less, on average [6]. However, the tradeoff with a more practical form is a reduction in precision. The SF-12 only uses one or two items to measure each of the eight SF-36 concepts. These scales have been shown to have significantly less precision than longer multi-item scales [17]. Thus, each of the eight health concepts is measured with less precision by the SF-12, relative to the SF-36, because the SF-12 scales define fewer scale levels and pool less reliable variance [6]. Accordingly, U.S. studies have shown that in clinical tests, the empirical validity of the SF-12 physical and mental summary scores typically has been about 10% below that of the SF-36 summary measures [6,7]. Similar results were seen in this study, in comparisons of the relative validity of the PCS-12 and PCS-36 to detect age differences in the nine European countries.

For large group comparisons and longitudinal monitoring, the differences in measurement reliability of the SF-12 and SF-36 are less important, because confidence intervals for group averages in health scores are largely determined by sample size [6]. Thus, if a study focuses on measuring overall physical and mental health outcomes rather than the eight-scale profile, and the sample size is large (n = 500+), the SF-12 may be advantageous. For smaller studies, and for studies in which the focus is on one or more of the eight SF-36 concepts rather than the two summary measures, the SF-36 is preferred. Research currently is ongoing in the United States to calibrate the eight-scale health profile across the SF-36 and SF-12, which will increase the usefulness of the SF-12 in smaller studies. Further research is needed to determine the extent to which the same tradeoffs between the SF-36 and SF-12 are involved in the nine countries studied here and in other IQOLA countries.

References

1. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. **Med Care** 1992; 30(6): 473–483.
2. Ware JE, Snow KK, Kosinski M, Gandek B. **SF-36 Health Survey Manual and Interpretation Guide**. Boston, MA: New England Medical Center, The Health Institute; 1993.
3. Ware JE, Kosinski M, Keller SK. **SF-36 Physical and Mental Health Summary Scales: A User's Manual**. Boston, MA: The Health Institute; 1994.
4. 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): 247–263.
5. Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek AE. Comparison of methods for 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.
6. Ware JE, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. **Med Care** 1996; 34(3): 220–233.
7. Ware JE, Kosinski M, Keller SD. **SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. Second Edition**. Boston, MA: The Health Institute, New England Medical Center; 1995.
8. Ware JE, Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P, *et al.* The factor structure of the SF-36 Health Survey in ten countries: Results from the IQOLA Project. **J Clin Epidemiol** 1998; 51(11): 1159–1165.
9. Ware JE, Gandek B, Kosinski M, Aaronson NK, Apolone G, Brazier J, *et al.* The equivalence of SF-36 summary health scores estimated using standard and country-specific algorithms in ten countries: Results from the IQOLA Project. **J Clin Epidemiol** 1998; 51(11): 1167–1170.
10. Gandek B, Ware JE. Methods for validating and norming translations of health status questionnaires: The IQOLA Project approach. **J Clin Epidemiol** 1998; 51(11): 953–959.
11. Sullivan M, Karlsson J, Ware JE. **SF-36 Hälsoenkät: Svensk Manual Och Tolkningsguide (Swedish Manual and Interpretation Guide)**. Gothenburg: Sahlgrenska University Hospital; 1994.
12. Apolone G, Mosconi P, Ware JE. **Questionario sullo stato di Salute SF-36. Manuale d'use e guida all interpretazione d'ei risultati**. Milano: Guerini & Associato Editore; 1997.
13. Gandek B, Ware JE, Aaronson NK, Alonso J, Apolone G, Bjorner JB, *et al.* Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: Results from the IQOLA Project. **J Clin Epidemiol** 1998; 51(11): 1149–1158.
14. Ware JE, Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. **J Clin Epidemiol** 1998; 51(11): 903–912.
15. Schofield MJ, Mishra G. Validity of the SF-12 compared with the SF-36 Health Survey in pilot studies of the Australian Longitudinal Study on Women's Health. **J Health Psychology** 1998; 3: 259–271.
16. Jenkinson C, Layte R. Development and testing of the UK SF-12. **J Health Serv Res Policy** 1997; 2: 14–18.
17. McHorney CA, Ware JE, Rogers WH, Raczek AE, Lu JFR. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: Results from the Medical Outcomes Study. **Med Care** 1992; 30(Suppl. 5) MS253–MS-265.