VALUE IN HEALTH 16 (2013) 1023–1031

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Preference-Based Assessments

Exploring the Consistency of the SF-6D

Lara N. Ferreira, PhD[1,2,], Pedro L. Ferreira, PhD[2,3], Luis N. Pereira, PhD[1,4], Donna Rowen, PhD[5], John E. Brazier, PhD[5]*

[1]*University of the Algarve-ESGHT, Faro, Portugal;* [2]*Centre for Health Studies & Research, University of Coimbra, Coimbra, Portugal;* [3]*Faculty of Economics, University of Coimbra, Coimbra, Portugal;* [4]*Research Centre for Spatial and Organizational Dynamics, University of the Algarve, Faro, Portugal;* [5]*Health Economics and Decision Science Section, School of Health and Related Research, University of Sheffield, Sheffield, UK*

ABSTRACT

Objective: The six-dimensional health state short form (SF-6D) was designed to be derived from the short-form 36 health survey (SF-36). The purpose of this research was to compare the SF-6D index values generated from the SF-36 ($SF\text{-}6D_{SF\text{-}36}$) with those obtained from the SF-6D administered as an independent instrument ($SF\text{-}6D_{Ind}$). The goal was to assess the consistency of respondents' answers to these two methods of deriving the SF-6D. **Methods:** Data were obtained from a sample of the Portuguese population (n = 414). Agreement between the instruments was assessed on the basis of a descriptive system and their indexes. The analysis of the descriptive system was performed by using a global consistency index and an identically classified index. Agreement was also explored by using correlation coefficients. Parametric tests were used to identify differences between the indexes. Regression models were estimated to understand the relationship between them. **Results:** The $SF\text{-}6D_{Ind}$ generates higher values than does the $SF\text{-}6D_{SF\text{-}36}$. There were significant differences between the indexes across sociodemographic groups. There was a significant ceiling effect in the $SF\text{-}6D_{Ind}$ but not in the $SF\text{-}6D_{SF\text{-}36}$. The correlation between the indexes was high but less than what was anticipated. The global consistency index identified the dimensions with larger differences. Considerable differences were found in two dimensions, possibly as a result of different item contexts. Further research is needed to fully understand the role of the different layouts and the length of the questionnaires in the respondents' answers. **Conclusions:** The results show that as the SF-6D was designed to derive utilities from the SF-36 it should be used in this way and not as an independent instrument.

Keywords: consistency, dimensions, SF-6D, SF-36.

Introduction

The short-form 36 health survey (SF-36) is a 36-item generic health status instrument, comprising eight scales and two component summary scales [1,2], which has been extensively validated and used e.g., [3–12]. The SF-36 is a profile-based patient-reported outcome measure that yields health scores across its eight dimensions. It does not, however, generate utilities, and hence it has a limited use in economic evaluations of health care interventions or technologies. To overcome this problem, a decade ago, Brazier et al. [13] developed an algorithm to translate the SF-36 results into health state utilities. They created the six-dimensional health state short form (SF-6D), an econometric preference-based index derived from 11 items of the SF-36, which are combined into six dimensions of health, with four to six levels each [13]. The SF-6D describes 18,000 different health states. A valuation survey was carried out in the United Kingdom to obtain values to a sample of 249 health states defined for the SF-6D. A representative sample of the general UK population valued these health states by using the standard gamble method. Econometric models were estimated by using the data collected to predict utility scores for all health states defined by the SF-6D [14]. These health state values constitute the SF-6D index, which can be seen as a continuous value ranging from 0.35 to 1.00. Another version of the SF-6D was developed on the basis of the short-form 12 health survey instrument (SF-12), and utility scores for all health states defined by this instrument for the UK population are also available [14]. The SF-6D enables a utility score to be generated by using responses to the SF-36 or the SF-12. There are now specific value sets for the SF-6D for Portugal [15], Japan [16], Hong Kong [17], and Brazil [18], with value sets for Australia and Singapore currently being determined.

Given that the SF-36 is widely used all over the world, the use of the SF-6D as a way of generating utilities from the SF-36 has increased in recent years and is now one of the preference-based indexes most widely used in cost-utility analyses and other studies that aim at measuring individuals' preferences for health states [e.g., 19–26] and included in numerous pharmacoeconomics guidelines. Previous research has focused on the assessment of the performance of the SF-6D and on comparisons with other

* *Address correspondence to*: Lara N. Ferreira, Escola Superior de Gestão, Hotelaria e Turismo, Universidade do Algarve, Campus da Penha, 8005-139 Faro, Portugal.

E-mail: Lnferrei@ualg.pt.

1098-3015/$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2013.06.018

preference-based indexes such as the EuroQol five-dimensional questionnaire or the Health Utilities Index. Several articles have been published on these topics [e.g., 27–34]. It is also useful to know whether the SF-6D health state classification can be used in its own right because this would be a more efficient way to collect the data. There are, however, no published studies having been dedicated to studying the consistency of the SF-6D when the classification system is used directly in a study to derive the health state of the individual rather than deriving the health state from responses to the SF-36 or the SF-12. Therefore, we intend to overcome this gap in the literature by exploring the consistency of respondents' answers to these two methods of deriving the SF-6D.

The aim of this research was twofold: 1) to test the hypothesis that the SF-6D applied as an independent instrument ($SF\text{-}6D_{Ind}$) produces results different from those obtained from the SF-6D index generated from the SF-36 ($SF\text{-}6D_{SF\text{-}36}$) and 2) to examine whether the conclusions differed depending on the value set used.

Methods

Sample and Data Collection

Data were collected from a sample of individuals from the adult general Portuguese (PT) population (n = 414) in spring 2011 in Portugal. Respondents were recruited from the population of students and staff of a public university in Portugal, according to their willingness to participate in the study. Although the sample used in the study is nonrandom, it was expected to include respondents from different sociodemographic groups given that the population of individuals comprised undergraduate and graduate students and teaching and nonteaching staff. In addition, it was not essential to use a random representative sample of the general population given that to achieve the aim of this research it was only necessary to prove that using the $SF\text{-}6D_{Ind}$ produces results different from those obtained from the $SF\text{-}6D_{SF\text{-}36}$ in at least one sample.

Respondents self-completed the SF-36v2 and the SF-6D on a voluntary and anonymous basis. This enables analysis of the consistency of respondents' answers to the two above-mentioned methods of deriving the SF-6D index; the SF-6D was also applied as an independent questionnaire.

The order of the self-completed paper-and-pencil questionnaires was fixed and was the same throughout the study: first, the SF-36, and second, the SF-6D classification system. In addition, respondents reported information on sociodemographic variables, such as sex, age, marital status, education, labor market participation, area of living, income, and the presence (or not) of a chronic disease.

The UK [14] and the PT [15] value sets for the SF-6D were both applied to the data collected to further examine whether the conclusions differed depending on the value set used. We have applied only these two value sets because there are no other European value set for the SF-6D and the UK value set is considered the gold standard. In fact, before the elicitation of the PT value set, studies conducted in Portugal used UK population values.

Statistical Analysis

Sample characteristics were first described by computing descriptive statistics for sociodemographic variables. The analysis of the degree of agreement between instruments was divided into two parts. First, an analysis based on the classification system of both instruments was performed, that is, an analysis of what respondents reported about their health in each instrument. This task started with a general descriptive analysis of the distribution of responses across dimensions in both instruments. Then, the degree of association between dimensions of the SF-6D was measured by using the Spearman's correlation coefficient. In addition, we used the following two measures based on square two-way contingency tables: a global consistency index (GCI) and an identically classified index (ICI). The GCI computes the percentage of individuals classified in the same level of each dimension in both instruments and is given by

$$GCI = \frac{\sum_{j=1}^{l} n_{jj}}{n} \times 100, \quad (1)$$

where n is the sample dimension and n_{jj} is the number of individuals with response in the same level j ($j = 1, \ldots, l$) of a particular dimension in the $SF\text{-}6D_{SF\text{-}36}$ and in the $SF\text{-}6D_{Ind}$. The GCI will be equal to 100 if all individuals equally respond on a specific dimension in both instruments. GCI values above 75 are interpreted as a strong agreement between instruments, whereas GCI values ranging from 50 to 75 are considered as a moderate agreement. GCI values lower than 50 suggest a poor agreement. The ICI calculates the percentage of individuals correctly classified in a level j of each dimension in the $SF\text{-}6D_{Ind}$ and is given by

$$ICI_j = \frac{n_{jj}}{n_{j\bullet}} \times 100, \quad (2)$$

where $n_{j\bullet} = \sum_{k=1}^{l} n_{jk}$ is the total number of responses in level j of a particular dimension in the $SF\text{-}6D_{SF\text{-}36}$. The ICI can be interpreted as a stability indicator and will be equal to 100 if all individuals equally respond on a level j of a specific dimension in both instruments. We also define a poor level of stability on responses when the ICI is less than or equal to 25.

Second, an analysis of the preference-based indexes generated by the instruments was carried out by using the following data analysis: 1) basic descriptive statistics including means, medians, and ranges to compare the main features of the indexes; 2) skewness statistics and one-sample Kolmogorov-Smirnov tests to evaluate the asymmetry and normality of distributions; 3) ceiling and floor effects (proportion of respondents with the best and worst possible theoretical scores, respectively) were identified; 4) Pearson's correlation coefficients to study the association between instruments and intraclass correlation coefficients (ICCs) based on a two-way mixed model with absolute agreement, for a global assessment of the agreement between indexes; 5) paired-samples t test (related-samples Wilcoxon signed-rank tests) to identify mean (median) differences between the indexes; and 6) regression analysis to explore the nature of their relationship. We have used the following model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad (3)$$

where Y represents the $SF\text{-}6D_{SF\text{-}36}$ index, X the $SF\text{-}6D_{Ind}$ index, ε the residuals, and i respondents ($i = 1, \ldots, n$). It should be noted, however, that the aim of the regression analysis was to test whether there is a perfect agreement between the indexes and not to explain or predict the $SF\text{-}6D_{SF\text{-}36}$ index through the $SF\text{-}6D_{Ind}$ index. Because an agreement between the indexes would result in estimated models in which the constant (α) would be equal to zero and the slope (β) equal to one, hypothesis tests were performed to verify these assumptions. Finally, the pattern of agreement was also examined graphically by plotting values obtained for the UK and PT value sets.

It should be noted that mean differences between indexes were evaluated by using paired-samples t tests, although the normality assumption was not verified. This decision was based on the following: the large sample size of our study; the well-known result that the power of the Kolmogorov-Smirnov Z test increases with the sample size; and some evidence that

nonparametric statistical methods produce similar results and the same conclusions to those of parametric methods, and that the latter are, thereby, robust to violation of assumptions such as normality [32]. Although we do not report detailed results, we have verified whether the conclusions were the same by using nonparametric tests. The purposes of the subgroup analyses performed were to understand whether there were significant differences between the subgroups and to identify the subgroups in which there were most likely to be the differences. All data analyses were performed by using the IBM SPSS Statistics version 19.0 package.

Results

Table 1 shows the main characteristics of the study sample. Age ranged from 16 to 70 years, with a mean age of 28 (SD=9.7). Of the respondents, 64.6% were women, 69.2% were single, and 54.5% had a middle education level. Furthermore, only 45.1% were actively employed, 35.4% earned less than €1000 per month, 79.3% lived in urban areas, and 77.5% did not report a chronic disease.

An inspection of the distributions of individuals' responses across dimensions of the two instruments (Table 2) revealed that there was a considerable percentage of responses in different levels of the same dimension in the SF-6D$_{Ind}$ and the SF-6D$_{SF-36}$. If it was acceptable to use the SF-6D as an independent instrument, we would expect respondents to report the same level in a specific dimension, whatever the instrument used; that is, distributions in two-way contingency tables would be near diagonal matrices.

Table 1 – Study sample characteristics.

Sociodemographic variables	n (%)
Sex	
Female	267 (64.6)
Male	146 (35.4)
Age group (y)	
≤20	92 (22.7)
21–40	264 (65.0)
>40	50 (12.3)
Marital status	
Single	285 (69.2)
Married/living together	113 (27.4)
Divorced/separated	10 (2.4)
Widowed	4 (1.0)
Educational level	
Low	14 (3.4)
Middle	225 (54.5)
High	174 (42.1)
Relation to the labor market	
Actively employed	185 (45.1)
Actively unemployed	12 (2.9)
Inactive	213 (52.0)
Area of living	
Urban area	326 (79.3)
Rural area	85 (20.7)
Income (€/month)	
<1000	142 (35.4)
1000–2000	153 (38.3)
>2000	105 (26.3)
Chronic disease	
Yes	89 (22.5)
No	307 (77.5)

For an in-depth analysis of the level of agreement, we have computed measures of agreement between instruments for each dimension, based on the respondents' self-reported health, which are presented in Table 3. In this table, we also present the distribution of responses in each dimension of the SF-6D$_{SF-36}$. Because results were the same whatever the value set used (PT or UK), we decided to not refer the value set in that table. Although it was observed that individuals responded differently on the same dimension of the instruments, one would expect to find strong direct correlations between the same dimension in the SF-6D$_{Ind}$ and the SF-6D$_{SF-36}$ because the Spearman's correlation coefficient assesses how well the relationship between two variables can be described but does not measure whether the rank values are the same in both variables. That was the case because all the correlations were statistically significant and some of them can be seen as moderate/strong. Indeed, the highest correlation of 0.667 was found between the responses in these instruments for pain, followed by the correlation between the responses for physical functioning ($\rho = 0.570$) and social functioning ($\rho = 0.566$). The lowest correlation was observed between the responses in these instruments for role limitations ($\rho = 0.412$).

An analysis of the results of the GCI reveals important differences in the level of agreement in responses across the dimensions of the SF-6D. For example, 69.2% of the individuals responded at the same level in the SF-6D$_{Ind}$ and the SF-6D$_{SF-36}$ for physical functioning [$GCI_{physical\ functioning} = (215 + 68 + 2 + 0 + 0 + 0) \times 100/412 = 69.2$]. In contrast, only 25.2% of the individuals responded at the same level for vitality and 28.2% for mental health. Although some of the GCI values can be seen as moderate (e.g., physical functioning, social functioning, and pain), none of them shows a strong agreement between the instruments.

This evidence is also supported by the results of the ICI. Agreement between responses in each level is far from perfect. Indeed, in level 1 of each dimension (no problems), there is good agreement because all ICI values were above 84% [e.g., $ICI_{1,\ physical\ functioning} = 215 \times 100/229 = 93.9$, $ICI_{1,\ role\ limitations} = 138 \times 100/146 = 94.5$]. Instability on responses, however, is evident in worst categories. For example, there is no agreement in the most severe levels (extreme problems) for physical functioning, social functioning, pain, and vitality, but the number of respondents who reported their health in these levels is too small to merit generalization. A poor level of stability on responses was observed in level 3 of physical functioning, levels 3 and 4 of role limitations, levels 3 and 5 of pain, and levels 3 and 4 of mental health and vitality in which there are larger numbers. This means that at least 75% of those reporting these specific levels in the SF-6D$_{SF-36}$ did not equally rate their health using the SF-6D$_{Ind}$.

Descriptive statistics of the SF-6D indexes are presented in Table 4. First, it should be noted that SF-6D$_{Ind}$ modal scores are equal to one using both the UK and PT value sets. The mean scores of SF-6D$_{SF-36}$ (PT) and SF-6D$_{Ind}$ (UK) are lower than the median scores, while the mean scores of SF-6D$_{Ind}$ (PT) and SF-6D$_{SF-36}$ (UK) slightly exceed the median scores. The Kolmogorov-Smirnov Z test showed non-normal data distribution for all four indexes. The skewness results, however, show that departures from symmetry are not remarkable, particularly for the SF-6D$_{SF-36}$ (UK).

Furthermore, the SF-6D$_{Ind}$ generated higher scores than did the SF-6D$_{SF-36}$ whatever the value set used (PT or UK). Indeed, all descriptive statistics of central tendency computed for SF-6D$_{Ind}$ utility scores clearly and systematically exceed those computed for the SF-6D$_{SF-36}$ utility scores (Table 4). These results were confirmed by the results of paired-samples t tests and related-samples Wilcoxon signed-rank tests. These tests confirm the existence of statistically significant mean and median differences, respectively, between indexes generated by the SF-6D$_{SF-36}$

Table 2 – Distributions of individuals' responses across the dimensions of the two instruments.

Physical functioning

$SF\text{-}6D_{SF\text{-}36}$	$SF\text{-}6D_{Ind}$						Σ
	1	2	3	4	5	6	
1	215	10	0	1	3	0	229
2	57	68	12	0	0	0	137
3	11	22	2	0	0	0	35
4	0	0	0	0	0	0	0
5	1	4	1	0	0	0	6
6	3	2	0	0	0	0	5
Σ	287	106	15	1	3	0	412

Pain

$SF\text{-}6D_{SF\text{-}36}$	$SF\text{-}6D_{Ind}$						Σ
	1	2	3	4	5	6	
1	116	5	2	0	0	0	123
2	23	51	14	3	1	0	92
3	23	56	30	11	2	0	122
4	8	13	13	14	4	0	52
5	0	1	3	10	3	0	17
6	1	0	0	0	1	0	2
Σ	171	126	62	38	11	0	408

Role limitations

$SF\text{-}6D_{SF\text{-}36}$	$SF\text{-}6D_{Ind}$				Σ
	1	2	3	4	
1	138	5	2	1	146
2	18	6	0	0	24
3	90	7	26	0	123
4	58	23	21	15	117
Σ	304	41	49	16	410

Mental health

$SF\text{-}6D_{SF\text{-}36}$	$SF\text{-}6D_{Ind}$					Σ
	1	2	3	4	5	
1	43	3	2	0	0	48
2	95	55	2	1	0	153
3	52	79	10	3	0	144
4	8	30	14	5	1	58
5	1	0	2	3	3	9
Σ	199	167	30	12	4	412

Social functioning

$SF\text{-}6D_{SF\text{-}36}$	$SF\text{-}6D_{Ind}$					Σ
	1	2	3	4	5	
1	189	16	3	0	0	208
2	67	43	7	1	0	118
3	19	21	16	0	0	56
4	3	8	5	11	0	27
5	2	0	0	1	0	3
Σ	280	88	31	13	0	412

Vitality

$SF\text{-}6D_{SF\text{-}36}$	$SF\text{-}6D_{Ind}$					Σ
	1	2	3	4	5	
1	22	4	0	0	0	26
2	100	61	3	1	0	165
3	38	74	12	5	0	129
4	13	37	16	8	1	75
5	1	1	6	6	0	14
Σ	174	177	37	20	1	409

SF-36, short-form 36 health survey; SF-6D, six-dimensional health state short form; $SF\text{-}6D_{SF\text{-}36}$, SF-6D index values generated from the SF-36; $SF\text{-}6D_{Ind}$, SF-6D index values generated from the SF-6D administered as an independent instrument.

Table 3 – Rank correlations, GCI, and ICI between instruments for each dimension.

Dimension	Spearman correlation	GCI	ICI (n)					
			1	2	3	4	5	6
Physical functioning	0.570*	69.2	93.9 (229)	49.6 (137)	5.7 (35)	ND (0)	0.0 (6)	0.0 (5)
Role limitations	0.412*	45.1	94.5 (146)	25.0 (24)	21.1 (123)	12.8 (117)	–	–
Social functioning	0.566*	62.9	90.9 (208)	36.4 (118)	28.6 (56)	40.7 (27)	0.0 (3)	–
Pain	0.667*	52.5	94.3 (123)	55.4 (92)	24.6 (122)	26.9 (52)	17.6 (17)	0.0 (2)
Mental health	0.499*	28.2	89.6 (48)	35.9 (153)	6.9 (144)	8.6 (58)	33.3 (9)	–
Vitality	0.504*	25.2	84.6 (26)	37.0 (165)	9.3 (129)	10.7 (75)	0.0 (14)	–

ND, not defined (there were no responses).
* $P < .001$.

and indexes generated by the SF-6D$_{Ind}$ using both the PT and UK value sets. In Table 4, we can also observe that mean differences between the SF-6D$_{Ind}$ and the SF-6D$_{SF-36}$ utility scores are stronger when the UK value set is used (PT: −0.066; UK: −0.121).

Similar conclusions were obtained by sociodemographic characteristics (Table 5), although few statistically significant differences between sociodemographic subgroups exist. Mean scores of all indexes were considerably higher for males than for females, and higher in the subgroup who did not report a chronic disease.

Table 4 shows that floor effects were not observed on all indexes. The SF-6D$_{Ind}$ indexes, however, showed a non-negligible ceiling effect, with 27.5% and 18.4% of the individuals having the highest possible score when using the PT and the UK value sets, respectively. This compares to a ceiling effect of 2.0% and 1.2%, respectively, for the SF-6D$_{SF-36}$ indexes. The ceiling effect of the SF-6D$_{Ind}$ indexes varies substantially across different sociodemographic groups, even though it presents approximately the same distribution for both value sets, which is expected given that the underlying health state is the same regardless of the value set applied (Table 6). For example, the ceiling effect for males is visibly larger than that for females; the ceiling effect in the middle education subgroup is slightly larger than that in the high education subgroup and is zero in the low education subgroup; the ceiling effect in the subgroup who reported chronic disease is considerably smaller than that in the opposing subgroup.

Over the study sample, the SF-6D$_{SF-36}$ and SF-6D$_{Ind}$ scores were correlated, although far from a perfect correlation, regardless of the value set used (PT: $r = 0.677$, ICC = 0.467; UK: $r = 0.709$, ICC = 0.473). While the correlation between the SF-6D$_{SF-36}$ and SF-6D$_{Ind}$ scores was higher when the UK value set was used for the overall sample, there were some subgroups in which we observe the opposite (Table 5). In Table 5, we can also observe that perfect correlations do not exist between indexes in any subgroup and the level of correlation varied according to sociodemographic characteristics.

The plot of the SF-6D$_{SF-36}$ to the SF-6D$_{Ind}$ (Fig. 1) shows a substantial deviation from the line that would be indicative of the perfect agreement between the two measures (i.e., the 45-degree line from the origin of 0.40–1), with a high percentage of observations below that line (88.1% and 87.8%, respectively, in the case of PT and UK value sets). The plots clearly demonstrate the ceiling effects in the SF-6D$_{Ind}$. These plots suggest that estimated linear regression models would not have constant equal to zero and slope equal to one. This conjecture was confirmed by the estimated coefficients, respectively, presented below each scatter plot of Figure 1 and by their P values associated with the following null hypothesis: $\alpha = 0$ and $\beta = 1$. All P values were less than .001 in both models. Thus, the coefficients indicate that the relationship differs under the value set used. Furthermore, both estimated models had a constant statistically significantly different to zero and slope different to one.

Discussion

This study aimed at exploring the consistency of the SF-6D by comparing the results of SF-6D$_{SF-36}$ with those of SF-6D$_{Ind}$. The PT and UK country-specific value sets were used to examine whether

Table 4 – Descriptive statistics of the indexes on the study sample.

	SF-6D$_{SF-36}$ (PT)	SF-6D$_{Ind}$ (PT)	SF-6D$_{SF-36}$ (UK)	SF-6D$_{Ind}$ (UK)
Observed range (theoretical: 0.35–1.00)	0.62–1.00	0.67–1.00	0.42–1.00	0.52–1.00
Mode	0.922	1.000	0.887	1.000
Mean (SD)	0.861 (0.075)	0.927 (0.066)	0.729 (0.117)	0.850 (0.124)
Median (IQR)	0.881 (0.81–0.92)	0.924 (0.88– 1.00)	0.714 (0.64–0.83)	0.884 (0.76–0.96)
KS Z test	0.128*	0.137*	0.090*	0.114*
Skewness (SE)	−0.843 (0.323)	−0.827 (0.325)	0.203 (0.123)	−0.595 (0.220)
Ceiling effect (%)	1.96	27.45	1.23	18.38
Mean difference†	−0.066*		−0.121*	
Median difference‡	−0.043*		−0.170*	
Pearson correlation	0.677*		0.709*	
ICC	0.467*		0.473*	

ICC, intraclass correlation coefficient; IQR, interquartile range; KS, Kolmogorov-Smirnov; PT, Portuguese; SE, standard error; SF-36, short-form 36 health survey; SF-6D, six-dimensional health state short form; SD, standard deviation; SF-6D$_{SF-36}$, SF-6D index values generated from the SF-36; SF-6D$_{Ind}$, SF-6D index values generated from the SF-6D administered as an independent instrument..
* $P < .001$.
† According to paired-samples t test.
‡ According to related-samples Wilcoxon signed-rank tests.

Table 5 – Index means, mean differences, and levels of correlation by sociodemographic characteristics.

Sociodemographic variables	PT value set					UK value set				
	$SF\text{-}6D_{SF\text{-}36}$	$SF\text{-}6D_{Ind}$	Mean difference*	Pearson correlation	ICC	$SF\text{-}6D_{SF\text{-}36}$	$SF\text{-}6D_{Ind}$	Mean difference*	Pearson correlation	ICC
Sex										
Female	0.851	0.915	−0.064†	0.696†	0.497†	0.709	0.828	−0.119†	0.719†	0.479†
Male	0.880	0.949	−0.069†	0.577†	0.346†	0.766	0.890	−0.124†	0.646†	0.404†
t test‡	−3.882†	−5.101†				−4.727†	−5.028†			
Age group (y)										
≤20	0.867	0.932	−0.065†	0.601†	0.428†	0.732	0.860	−0.128†	0.679†	0.436†
21–40	0.859	0.928	−0.068†	0.674†	0.440†	0.727	0.852	−0.124†	0.704†	0.450†
>40	0.857	0.912	−0.055†	0.796†	0.635†	0.732	0.822	−0.090†	0.780†	0.642†
F test§	0.741	1.712				0.168	1.580			
Marital status										
Single	0.861	0.929	−0.068†	0.641†	0.432†	0.726	0.851	−0.125†	0.682†	0.441†
Married/living together	0.861	0.923	−0.061†	0.745†	0.543†	0.737	0.850	−0.113†	0.765†	0.544†
Divorced/separated	0.879	0.942	−0.063†	0.902†	0.602†	0.747	0.873	−0.126†	0.886†	0.596†
Widowed	0.830	0.897	−0.067	0.829	0.636‖	0.681	0.751	−0.070	0.874	0.543
F test§	0.418	0.680				0.447	0.977			
Educational level										
Low	0.809	0.878	−0.069†	0.858†	0.607†	0.661	0.765	−0.104†	0.836†	0.626†
Middle	0.861	0.931	−0.070†	0.622†	0.416†	0.730	0.859	−0.129†	0.677†	0.433†
High	0.866	0.926	−0.060†	0.731†	0.518†	0.733	0.845	−0.112†	0.738†	0.510†
F test§	3.615‖	4.404‖				2.391	4.027‖			
Chronic disease										
Yes	0.831	0.888	−0.057†	0.798†	0.610†	0.696	0.783	−0.087†	0.738†	0.630†
No	0.868	0.937	−0.069†	0.610†	0.394†	0.738	0.868	−0.130†	0.675†	0.416†
t test‡	−4.430†	−6.428†				−3.190*	−5.848†			

ANOVA, analysis of variance; ICC, intraclass correlation coefficient; PT, Portuguese; SF-36, short-form 36 health survey; SF-6D, six-dimensional health state short form; $SF\text{-}6D_{SF\text{-}36}$, SF-6D index values generated from the SF-36; $SF\text{-}6D_{Ind}$, SF-6D index values generated from the SF-6D administered as an independent instrument.

* According to paired-samples *t* test, but the conclusions were the same if related-samples Wilcoxon signed-rank tests were used.

† $P < .001$.

‡ According to independent-samples *t* test, but the conclusions were the same if independent-samples Mann-Whitney *U* tests were used.

§ According to one-way ANOVA, but the conclusions were the same if independent-samples Kruskal-Wallis tests were used.

‖ $P < .05$.

Table 6 – Ceiling effect (%) by sociodemographic characteristics.

Sociodemographic variables	PT		UK	
	SF-6D$_{SF-36}$	SF-6D$_{Ind}$	SF-6D$_{SF-36}$	SF-6D$_{Ind}$
Sex				
Female	0.8	20.1	0.4	12.5
Male	4.2	41.0	2.8	29.2
Age group (y)				
≤20	5.4	34.1	3.3	16.5
21–40	0.8	26.4	0.8	20.3
>40	0.0	18.4	0.0	8.2
Marital status				
Single	2.8	29.1	1.8	19.5
Married/living together	0.0	22.5	0.0	15.3
Divorced/separated	0.0	40.0	0.0	30.0
Widowed	0.0	25.0	0.0	0.0
Educational level				
Low	0.0	0.0	0.0	0.0
Middle	2.7	29.1	1.8	19.3
High	1.2	27.5	0.6	18.7
Chronic disease				
Yes	0.0	9.3	0.0	7.0
No	2.3	32.8	1.3	21.3

PT, Portuguese; SF-36, short-form 36 health survey; SF-6D, six-dimensional health state short form; SF-6D$_{SF-36}$, SF-6D index values generated from the SF-36; SF-6D$_{Ind}$, SF-6D index values generated from the SF-6D administered as an independent instrument.

the conclusions differed depending on the value set used. The SF-6D$_{Ind}$ generates higher values than does the SF-6D$_{SF-36}$ using both country value sets. There were also significant differences between the indexes across sociodemographic groups.

We found a significant ceiling effect in the SF-6D$_{Ind}$ not observed in the SF-6D$_{SF-36}$. Moreover, this ceiling effect of the SF-6D$_{Ind}$ index varied substantially across different sociodemographic groups. If this variability just reflected the fact that men, young, those with a high income, and those without chronic disease are healthier, this would be expected; however, the same results should also be expected in the SF-6D$_{SF-36}$, but this was not the case.

As mentioned before, many items in the SF-36 are used to generate the SF-6D$_{SF-36}$ whereas each SF-6D$_{Ind}$ dimension is generated by using only one item. Our results showed that the use of the SF-6D descriptive system directly led the individuals to answer on the higher level. To understand this finding, we may speculate that a larger context of questions, that is, a larger range for the concept to be measured, may change the relative location

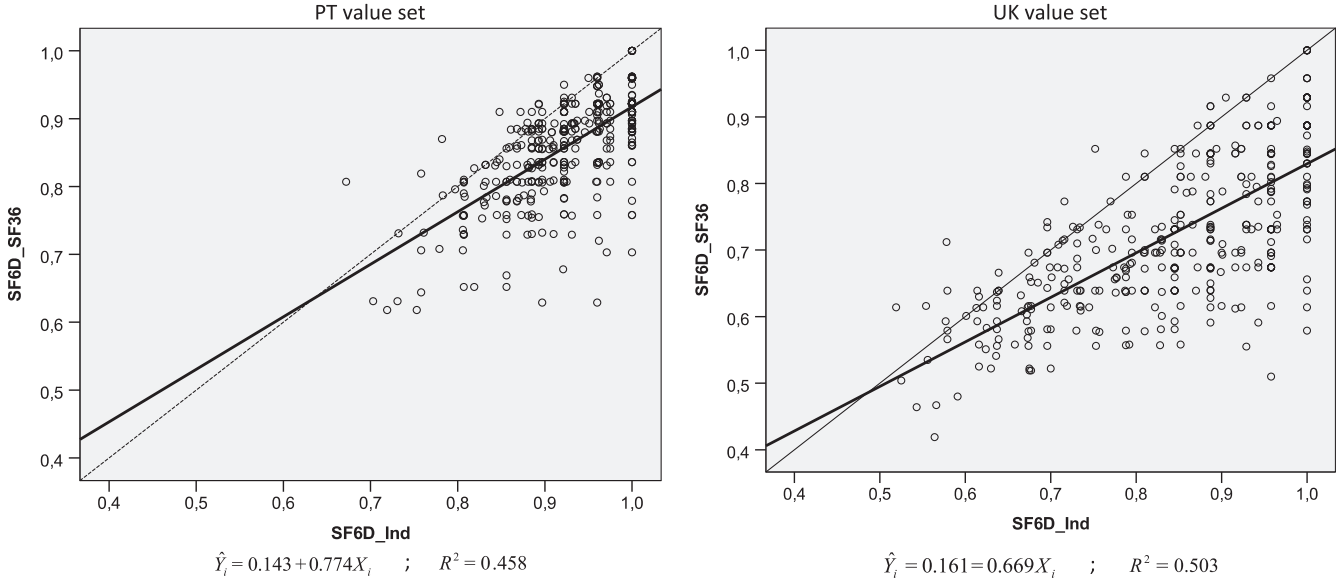

Fig. 1 – Relationship between the distribution of the SF-6D$_{SF-36}$ and the SF-6D$_{Ind}$ utility scores (whole sample). SF-36, short-form 36 health survey; SF-6D, six-dimensional health state short form; SF-6D$_{SF-36}$, SF-6D index values generated from the SF-36; SF-6D$_{Ind}$, SF-6D index values generated from the SF-6D administered as an independent instrument.

of a particular question and so, the stimulus given to the respondent.

Furthermore, the ordering of the questionnaires in the survey meant that respondents always completed the SF-36 first, meaning by the time they reached the SF-6D classification system they had already completed many similar questions about their health. There are few studies in the literature that address issues related to order effects. A recent study [35], however, showed a tendency to not use the "in-between" levels when a longer version of a questionnaire was first scored than the original version. Therefore, in our opinion, changing the order of the instruments could have introduced "noise" in the study and probably influenced the results.

The results show that the SF-6D should be derived from the SF-36/SF-12 responses and should not be used as an independent instrument where respondents are asked to complete the classifications system alone because this will lead to different responses and different utility values. The findings provide evidence that moving from the $SF\text{-}6D_{SF\text{-}36}$ to the $SF\text{-}6D_{Ind}$ involves differences in the mean index values across all sociodemographic groups.

Although these results should not be generalized to other instruments, similar results could have been found if the Health Utilities Index mark 2 and 3 have been used because the latter have been adapted from the first. Arguably, these results could also apply to any other instrument in which a health state is derived from a longer instrument, including a number of condition-specific measures, such as the Asthma Quality of Life Questionnaire, the Overactive Bladder Questionnaire, or the European Organisation for Research and Treatment of Cancer questionnaire, from which preference-based single indexes have been derived. We hope that the methodology presented here can be of assistance in further studies on these issues.

The differences found constitute an important and worrying result, and the reasons should be further investigated. But why should we expect no differences? Overall, there is no reason why the instruments should give the same results. Some random differences would be expected because of measurement error—people simply make mistakes at each administration, but this should not result in any overall differences. A greater error might be expected in well-being items than in functioning, and this would be translated into greater differences. This may explain why mental health and vitality have lower agreement, but it is not a reason for bias. Similar to this is that the well-being items—mental health and vitality— responses are more evenly distributed (i.e., far fewer at 1.0) and so there is more scope for error. Those with large numbers at level 1 are bound to show higher levels of agreement. Again, this is not a reason for bias but may explain differences in the GCI. It could also be expected that those dimensions made up one item—such as vitality—to have more error because those composed of two or three SF-36 items could be more reliable.

Other reasons that could explain main differences concern the way respondents dealt with end points of the scales, the number of items used, and the context. In truth, the literature suggests that respondents avoid the ends of a scale (end-point bias); however, this was not the case for these respondents. In cases in which more than one item is used, there is more chance of indicating a problem in the SF-36 than in the SF-6D. This does not, however, explain the result for vitality.

The most likely explanation, however, is context. Our responses to items are altered by the context—in this case, the other items being completed, particularly those completed just before. In this case, it may be that completing items in the context of other related items—such as the 10 physical functioning items or the well-being items for mental health and vitality—makes us more likely to own up to problems we may have. Although when we complete one per dimension, it is more of an overview and so we tend to generalize to "no problem." One way to try to fully understand this would be to 1) repeat in a patient group; 2) repeat for other instruments, such as the condition-specific ones, such as the Asthma Quality of Life Questionnaire or the Health Utilities Index mark 3; and 3) undertake some qualitative work.

The overall implications of these findings may extend beyond the SF-6D to any short-form version of an instrument. Recently, there has been a big push to develop short-form instruments based on subsets of items derived by using Rasch or item response theory (e.g., patient-reported outcomes), and this research suggests that this may have implications for how items are completed.

The sample used (nonrandom sample) and its specific characteristics that mean it is not representative of the PT population (e.g., youth) could be considered as a limitation of this study. Arguably, however, this does not constitute a real drawback of the study. We strongly believe that although women and young people are overweighed in the sample, this does not have a significant impact on the conclusions of the study given the aim of this study. In fact, it is sufficient to show in a particular case that the SF-6D cannot be applied independently to prove that it is not valid in all cases.

The study, however, has a limitation that should be mentioned: the health states observed. Given that the sample was relatively young, the number of severe health states was low. The findings showed no agreement in the most severe levels (extreme problems) for physical functioning, social functioning, pain, and vitality. Because the number of respondents who reported their health in these levels was low, it was not possible to generalize the conclusions. Ideally, this would be done with a better distribution of ratings. Therefore, further research is encouraged to determine whether the same relationship between the indexes is observed for a sicker patient population.

Further research is needed to understand why respondents did not give similar answers when the SF-6D classification system was used directly and to fully understand the role of the different layouts and the length of the questionnaires in the respondents' answers. Qualitative research is recommended. For instance, following completion of the instruments, respondents could be shown their answers and asked to consider the reasons for any discrepancies. It would also be useful to implement a talk-aloud protocol to better understand the meaning given by the respondents to the wording of the instruments.

Conclusions

The results of this study provide evidence that the SF-6D should not be used as an independent instrument. It was designed to derive utilities from the SF-36 (or the SF-12) and should be used in this way.

Acknowledgments

We thank the editor and the anonymous reviewers for their constructive comments and suggestions, which have considerably improved an earlier version of the article. We also thank the Fundação para a Ciência e a Tecnologia (FCT) for financing the following research centers: Centre for Health Studies & Research-University of Coimbra and Research Centre for Spatial and Organizational Dynamics-University of the Algarve.

Sources of financial support: The authors have no other financial relationships to disclose.

REFERENCES

[1] Ware J, Sherbourne C. The MOS 36-item Short-Form Health Survey (SF-36): conceptual framework and item selection. Med Care 1992;30:473–83.
[2] Ware J, Snow K, Kosinski M, Gandek B. SF-36 Health Survey Manual and Interpretation Guide. Boston, MA: The Health Institute, 1993.
[3] Anderson C, Laubscher S, Burns R. Validation of the Short Form 36 (SF-36) Health Survey Questionnaire among stroke patients. Stroke 1996;27:1812–6.
[4] Bousquet J, Knani J, Dhivert H, et al. Quality of life in asthma, I: internal consistency and validity of the SF-36 questionnaire. Am J Resp Crit Care Med 1994;149:371–5.
[5] Hawthorne G, Osborne R, Taylor A, Sansoni J. The SF36 Version 2: critical analyses of population weights, scoring algorithms and population norms. Qual Life Res 2007;16:661–73.
[6] Hopman W, Towheed T, Anasassiades T, et al. Canadian normative data for the SF-36 health survey. Can Med Assoc J 2000;168:265–71.
[7] Hsiung PC, Fang CT, Chang YY, et al. Comparison of WHOQOL-BREF and SF-36 in patients with HIV infection. Qual Life Res 2005;14:141–50.
[8] Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age. BMJ 1993;29:1437–40.
[9] Loge J, Kaasa S. Short Form 36 (SF-36) health survey: normative data from the general Norwegian population. Scand J Pub Health 1998;26:250–8.
[10] Mahler D, Mackowiak J. Evaluation of the Short-Form 36-Item Questionnaire to measure health-related quality of life in patients with COPD. Chest 1995;107:1585–9.
[11] Mosconi P, Cifani S, Crispino S, et al. The performance of SF-36 health survey in patients with laryngeal cancer head and neck cancer Italian working group. Head Neck 2000;22:175–82.
[12] Picavet H, Hoeymans N. Health related quality of life in multiple musculoskeletal diseases: SF-36 and EQ-5D in the DMC3 study. Ann Rheum Dis 2004;63:723–9.
[13] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271–92.
[14] Brazier J, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care 2004;42:851–9.
[15] Ferreira L, Ferreira P, Pereira L, et al. A Portuguese value set for the SF-6D. Value Health 2010;13:624–30.
[16] Brazier J, Fukahara S, Roberts J, et al. Estimating a preference-based index from the Japanese SF-36. J Clin Epidemiol 2009;62:1323–31.
[17] Lam C, Brazier J, McGhee S. Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a Chinese population. Value Health 2008;11:295–303.
[18] Cruz L, Camey S, Hoffmann J, et al. Estimating the SF-6D value set for a population-based sample of Brazilians. Value Health 2011;14(Suppl.):S108–14.
[19] Crawford B, Brazier J, Strand V, Doyle J. Treatment with leflunomide improves the utility of patients with active rheumatoid arthritis: an application of the SF-6D. Value Health 2001;4:70–1.
[20] Ferreira L, Ferreira P, Pereira L, Brazier J. An application of the SF-6D to create health values in Portuguese working age adults. J Med Econ 2008;11:215–33.
[21] Ferreira L, Ferreira P, Baleiro R. Quality of life in patients with rheumatoid arthritis. Acta Reumatológica Portuguesa 2008;33:341–2. [article in Portuguese].
[22] Ferreira L, Brito U, Ferreira P. Quality of life in asthma patients. Rev Port Pneumol 2010;XVI:23–55.
[23] Kortt M, Clarke P. Estimating utilities values for health states of overweight and obese individuals using the SF-36. Qual Life Res 2005;14:2177–85.
[24] Lee B, King M, Simpson J, et al. Validity, responsiveness, and minimal important difference for the SF-6D health utility scale in a spinal cord injured population. Value Health 2088;11:680-8.
[25] Sach T, Barton G, Jenkinson C, et al. Comparing cost-utility estimates: does the choice of EQ-5D or SF-6D matter? Med Care 2009;47:889–4.
[26] Stevenson MD, Scope A, Sutcliffe PA. The cost-effectiveness of group cognitive behavioral therapy compared with routine primary care for women with postnatal depression in the UK. Value Health 2010;13:580–4.
[27] Barton G, Sach T, Avery A, et al. Comparing the performance of the EQ-5D and SF-6D when measuring the benefits of alleviating knee pain. Cost Eff Resour Alloc 2009;7:12.
[28] Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13:873–84.
[29] Cunillera O, Tresserras R, Rajmil L, et al. Discriminative capacity of the EQ-5D, SF-6D, and SF-12 as measures of health status in population health survey. Qual Life Res 2010;19:853–64.
[30] Ferreira P, Ferreira L, Pereira L. How consistent are health utility values? Qual Life Res 2008;17:1031–42.
[31] Konerding U, Moock J, Kohlmann T. The classification systems of the EQ-5D, the HUI II and the SF-6D: what do they have in common? Qual Life Res 2009;18:1249–61.
[32] Kontodimopoulos N, Pappa E, Papadopoulos A, et al. Comparing SF-6D and EQ-5D utilities across groups differing in health status. Qual Life Res 2009;18:87–97.
[33] Marra C, Woolcott J, Kopec J, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60:1571–82.
[34] Pickard A, Simon J, Jeffrey A, Feeny D. Responsiveness of generic health-related quality of life in stroke. Qual Life Res 2005;14:207–19.
[35] Janssen M, Birnie E, Haagsma J, Bonsel G. Comparing the standard EQ-5D three-level system with a five-level version. Value Health 2008;11:275–84.