Article:

Poder, T.G., Fauteux, V., He, J. et al. (1 more author) (2019) Consistency between three different ways of administering the Short Form 6 Dimension version 2. Value in Health, 22 (7). pp. 837-842. ISSN 1098-3015

https://doi.org/10.1016/j.jval.2018.12.012

Consistency between three different ways of administering the SF-6Dv2

Abstract

Objective: The Short Form 6 Dimension (SF-6D) is a multi-attribute utility instrument derived from the SF-36v2 quality of life questionnaire and is used to calculate QALYs (Quality Adjusted Life Years) on a scale 0 to 1. The SF-6Dv2 is a new version of the SF-6D. The aim of this study was to assess the consistency of respondents' answers to three different methods to administer this new version. **Methods**: SF-6Dv2 utility values were generated from the SF-36v2 using: 1) full questionnaire with 36 items (SF-6Dv2$_{\text{SF-36}}$); 2) subset questionnaire with 10 items (SF-6Dv2$_{\text{ind-10}}$); 3) SF-6Dv2 administered as an independent instrument (rephrased questionnaire with only 6 items (SF-6Dv2$_{\text{ind-6}}$)). The order of the three instruments was randomly allocated between respondents. **Results**: A total of 782 respondents from Quebec, Canada, were interviewed, out of which 697 fully completed the survey. Very few deviations in respondents' answers were observed between the three instruments, with mean weighted kappa of 0.79 (range 0.61-0.91) and mean global consistency index of 70% (range 54-83). Maximal difference in utility values generated was found between SF-6Dv2$_{\text{ind-10}}$ and SF-6Dv2$_{\text{ind-6}}$ (mean difference 0.016, p<0.01) while minimal difference was found between SF-6Dv2$_{\text{SF-36}}$ and SF-6Dv2$_{\text{ind-10}}$ (0.002, p=0.38). No ceiling effect was observed. **Conclusions**: The SF-6Dv2 was designed to derive utilities from the SF-36v2 and our results indicate that it is still preferable to use the full questionnaire although the difference with other variants of the questionnaire is very small. To use the SF-6Dv2 as an independent instrument will thus introduce minimal bias in utility values generated.

Key words: SF-6Dv2; consistency; health utility; Quebec.

Introduction

A quality adjusted life year (QALY) value is a measure of health related quality of life (HRQoL) that is used to guide decisions pertaining to allocation of health-care resources. A QALY can be generated in two different ways. It can be generated directly using elicitation techniques (e.g., standard gamble, time trade-off), or indirectly using multi-attribute utility instrument (MAUI) [1-3]. While direct elicitation techniques allow generating accurate measures of QALY for a specific population, this is very time consuming and not always appropriate when conducting a clinical study, especially in paediatric populations [4]. In contrast, a MAUI is easy to use as it comes with a pre-validated questionnaire that incorporates a multi-attribute classification system. In this system, a pre-determined utility weight can be assigned to each health status. These weights are typically elicited from a sample of the general population. The three most used MAUI are the EuroQol 5 dimensions (EQ-5D), the Short-Form 6 Dimensions (SF-6D), and the Health utility index (HUI) [2]. Unlike the two other MAUI, the SF-6D is derived from a HRQoL questionnaire, namely the Short-form 36 health survey version 2 (SF-36v2). The SF-36v2 is a 36-item generic health status instrument and is one of the most widely used HRQoL. However, the SF-36v2 is not adapted to generate utilities to calculate QALYs [5]. In order to convert its responses into QALY, Brazier et al. [6,7] developed a MAUI, the SF-6Dv1 (previously named SF-6D), using 11 items in the SF-36v2 combined in 6 health dimensions. These six dimensions, with 4 to 6 levels each, describe 18,000 different health states. Up to eight valuation surveys were carried out in different countries around the world to generate value sets that can be used to convert the SF-6Dv1 responses into utility values for QALY [8]. The first of these value set was generated in the United Kingdom [7]. Recently, an improved version of the SF-6D has been developed using classical psychometric, Rasch and Item Response Theory techniques [9]. This new version (SF-6Dv2) contains only 10 items from the SF-36v2 and can potentially be used in three different ways. First, like the SF-6Dv1, it can be used in combination with the responses to the full SF-36v2 questionnaire (SF-6Dv2$_{\text{SF-36}}$). Second, it may be used as an independent instrument using only the 10 items requested from the SF-36v2 (SF-6Dv2$_{\text{ind-10}}$). Third, since the 10 items from the SF-36v2 have been combined and rephrased to generate a questionnaire with 6 questions, the SF-6Dv2 may be used as an independent instrument with only these 6 questions (SF-6Dv2$_{\text{ind-6}}$). Compared with the SF-6Dv1, the SF-6Dv2 have the same 6 dimensions, but with 5 to 6 levels each, yielding up to 18,750 health states.

Conventionally, it has been recommended to use the SF-6D conjointly with the SF-36v2 [7]. This was confirmed by a study carried out by Ferreira et al. [10]. In this study, they tested if the SF-6Dv1 can be used as an independent instrument using only the 6 rephrased items from the SF-36v2 (SF-6Dv1$_{\text{ind-6}}$) (i.e., the classification system of the SF-6Dv1) or if it must be used along with the entire SF-36v2 questionnaire (SF-6Dv1$_{\text{SF-36}}$). To do so, the SF-6Dv1$_{\text{SF-36}}$ and SF-6Dv1$_{\text{ind-6}}$ were administered to a sample group of 414 respondents from the general Portuguese population. The researchers found that the use of the SF-6Dv1$_{\text{ind-6}}$ systematically generated higher values than that of the SF-6Dv1$_{\text{SF-36}}$. They also found a significant ceiling effect in the SF-6Dv1$_{\text{ind-6}}$ but not in the SF-6Dv1$_{\text{SF-36}}$. This led Ferreira et al. [10] to conclude that the SF-6Dv1 should not be used as an independent instrument. However, in their study, they systematically asked respondents to complete the SF-36v2 first and then the SF-6Dv1 as an independent instrument. This may have generated a bias towards the non-use of the SF-6Dv1 as an independent instrument since it was always administered secondly. Indeed, respondents may have been upset, annoyed or simply tired of answering the same questions a second time, which could have biased their answers (i.e., a repetition bias). In addition, the sample used in this study was not representative of the general population of Portugal, but comprised mostly of students and university staff, which may potentially have influenced the results by overestimating index values in a cohort in better health.

In the present study, we propose to test if the different formats in which the SF-6Dv2 can be used provide consistent results and if they can be used interchangeably. However, unlike in the study by Ferreira et al. [10], we randomly assigned the order of the instruments to avoid any ordering effect. In addition, we tested the SF-6Dv2. Since the SF-6Dv2 can potentially be used in three ways (i.e., with the full SF-36v2, with the 10 items alone, or with these items combined in a set of 6 questions), we performed three comparisons to evaluate each of the scenario: SF-6Dv2$_{\text{SF-36}}$ vs. SF-6Dv2$_{\text{ind-10}}$, SF-6Dv2$_{\text{SF-36}}$ vs. SF-6Dv2$_{\text{ind-6}}$ and SF-6Dv2$_{\text{ind-10}}$ vs SF-6Dv2$_{\text{ind-6}}$.

Methods

We conducted an internet survey in the Province of Quebec, Canada, in 2016. Respondents were solicited from a panel of 4,800 emails obtained from previous studies conducted by the authors in the general population and where respondents had voluntarily provided their emails for future research. Respondents were contacted in April and a reminder was sent in May. Respondents completed the SF-6Dv2$_{\text{SF-36}}$, SF-6Dv2$_{\text{ind-10}}$ and SF-6Dv2$_{\text{ind-6}}$ in a random order. The survey was voluntary and anonymous. In addition to the three SF-6Dv2 questionnaires, sociodemographic data were also collected. This study was approved by our ethics committee.

Since a value set is neither available for Quebec nor for the SF-6Dv2 [11], we used the value set model 10 developed for the United Kingdom (UK) by Brazier et al. [7] to compare the utility values obtained by the three instruments tested. Model 10 is the one recommended by Pickard et al. [12]. Considering that this value set has been elicited from the SF-6Dv1 with 11 items associated to 4 to 6 levels for each dimension, we made some changes to fit the set adequately with the SF-6Dv2 that has six questions. Specifically, the coefficient associated with level 5 in the physical functioning dimension was removed (i.e., this specific level was removed from the SF-6Dv2 classification system) and a coefficient was added for a fifth level in the role limitations dimension (i.e., a fifth level has been added in the SF-6Dv2 to consider a permanent role limitation; the coefficient corresponds to the highest value obtained in this dimension, that is -0.055). In addition, to maintain a good correspondence between the answers provided to the 10 items from the SF-36v2 and the answers to the rephrased 6 questions of the SF-6Dv2, it was necessary to recode some answers. Specifically, the physical functioning dimension was recoded so as to make it comply with the new structure of the SF-6Dv2 (i.e., answers from the SF-36v2 that do not have a correspondence in the SF-6Dv2 classification system take a value of 2) and the role limitation and mental health dimensions considered the worst answers to the two items in the SF-36v2 they are related to. The structure of each instrument derived from the SF-36v2 is provided in Table 1. The classification of the SF-6Dv2 is also provided in Table A1 in appendix.

Table 1. Characteristics of the SF-6D instrument

Dimension	Items extracted from the SF-36		Recoding SF-6Dv2$_{\text{ind-6}}$ from the SF-36
	SF-6Dv1$_{\text{SF-36}}$	SF-6Dv2$_{\text{SF-36}}$	
Physical functioning	3a, 3b and 3j	3a, 3b and 3j	2 if no correspondence
Role limitations	4c and 5b	4b and 5b	Worst answer
Social functioning	10	10	NA
Pain	7 and 8	7	NA
Mental health	9b and 9f	9b and 9f	Worst answer
Vitality	9e	9g	NA
# of items	11	10	6

Notes: NA is for not applicable

Statistical analysis

Descriptive statistics for sociodemographic variables were calculated. An analysis of the degree of agreement between instruments was performed in two steps. In the first step, we analysed the distribution of answers for each instrument and calculated the global consistency index (GCI), the identically classified index (ICI), Spearman's correlation coefficient and quadratic weighted Kappa. The GCI calculates the percentage of individuals classified in the same level of each dimension in the two instruments compared:

$$GCI= j=1lnjjGCI = \frac{\sum_{j=1}^{l} n_{jj}}{n} \times 100$$

(1)

where n is the total number of subject and n_{jj} is the number of individuals with response in the same level j (j = 1, …, l) of a particular dimension.

The ICI calculates the percentage of individuals correctly classified in a given level of each dimension in the comparison instrument (i.e. the addition of the main diagonal divided by the total of subjects):

$$ICI = \frac{n_{ij}}{n_{j\bullet}} \times 100$$

(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 comparison instrument.

In the second step, utility values were calculated using the value set model 10 of Brazier et al. [7] for each individual and each instrument and comparisons were conducted. These comparisons were performed for the following measurements: mean, median, range, Kurtosis, Skewness, ceiling and floor effect, Pearson's correlation coefficient, intra-class correlation coefficient (ICC), R squared and graphical plot distribution. When appropriate, differences were assessed with pairedsamples t-test or F test. P-values < 0.05 were considered significant. All analyses were conducted with Stata SE (StataCorp LP, Texas, USA).

Results

There was 782 subjects who started the questionnaire and 697 surveys were completed in full. The sample was well-balanced as regards to sex, age, marital status, labor market and area of living (Table 2). However, the sample included people who are more educated and affluent than the general population.

Table 2. Sociodemographic variables in our sample as compared to the province of Quebec

	N sample	% sample	N Quebec	% Quebec
Sex				
Female	401	57.5%	3,434,946	50.6%
Male	296	42.5%	3,351,217	49.4%
Age group				
≤ 20 years	4	0.6%	280,076	4.1%
21-40 years	147	21.1%	2,228,387	32.8%
41-60 years	317	45.5%	2,334,060	34.4%
> 60 years	229	32.9%	1,943,640	28.6%
Marital status				
Single	155	22.2%	1,801,686	26.5%
Married/living together	435	62.4%	3,918,129	57.7%
Divorced/separated	85	12.2%	617,130	9.1%
Widowed	22	3.2%	450,975	6.6%
Educational level				
Low	134	19.2%	2,828,135	48.5%
Middle	254	36.4%	1,981,400	34.0%
High	309	44.3%	1,002,285	17.2%
Labor market				
Employed	367	52.7%	4,097,000	59.9%
Student	21	3.0%	-	-
Retired	232	33.3%	-	-
Unemployed	77	11.0%	337,200	4.9%
Area of living				
Urban area	528	75.8%	6,368,270	81.0%
Rural area	169	24.2%	1,534,731	19.0%
Income (K$)				
< 25	103	14.8%	2,787,149	43.5%
25-50	152	21.8%	1,899,596	29.6%
50-70	135	19.4%	823,432	12.8%
70-100	161	23.1%	540,712	8.4%
≥ 100	146	20.9%	360,481	5.6%
HRQoL problem				
Yes	303	43.5%	-	-
No	394	56.5%	-	-

Notes: Data for Quebec are the latest available (Institut de la statistique du Québec, Statistique Canada, Revenu Québec). Sex, age group, and marital status data for Quebec pertain to people aged 18 or above, whereas other sociodemographic data are for people aged 15 or above.

The distribution of the order of administration of the SF-6Dv2 in our survey is presented in Table A2 in appendix. Since the distribution of the survey was at random, we have about the same number of respondents in each sequence (i.e., 1/6). In Table A3 in appendix, individual responses have been presented across the 6 dimensions and various levels of the three instruments compared. We found that most of the responses were located close to the diagonal, indicating that responses to each instrument were quite similar. Table 3 presents the deviations in responses between the 3 instruments. A deviation of 0 indicates that responses are the same and a deviation of 1 indicates that they only differ by one level. On average, there was no deviation in 70% of cases (range 54-83%), and more than 90% of responses differed by only one level or less.

Table 3. Deviation between responses in SF-6Dv2$_{SF-36}$, SF-6Dv2$_{ind-10}$ and SF-6Dv2$_{ind-6}$

	0	1	2	>2	Σ
SF-6Dv2SF-36 and SF-6Dv2ind-10					
Physical functioning	562 (80.6%)	116 (16.6%)	15 (2.2%)	4 (0.6%)	697
Role limitations	420 (60.3%)	242 (34.7%)	31 (4.4%)	4 (0.6%)	697
Social functioning	527 (75.6%)	154 (22.1%)	13 (1.9%)	3 (0.4%)	697
Pain	572 (82.1%)	114 (16.4%)	8 (1.1%)	3 (0.4%)	697
Mental health	465 (66.7%)	210 (30.1%)	19 (2.7%)	3 (0.4%)	697
Vitality	478 (68.6%)	197 (28.3%)	18 (2.6%)	4 (0.6%)	697
SF-6Dv2SF-36 and SF-6Dv2ind-6					
Physical functioning	498 (71.4%)	171 (24.5%)	20 (2.9%)	8 (1.1%)	697
Role limitations	407 (58.4%)	238 (34.1%)	47 (6.7%)	5 (0.7%)	697
Social functioning	519 (74.5%)	158 (22.7%)	18 (2.6%)	2 (0.3%)	697
Pain	579 (83.1%)	106 (15.2%)	6 (0.9%)	6 (0.9%)	697
Mental health	377 (54.1%)	260 (37.3%)	51 (7.3%)	9 (1.3%)	697
Vitality	456 (65.4%)	209 (30.0%)	26 (3.7%)	6 (0.9%)	697
SF-6Dv2ind-10 and SF-6Dv2ind-6					
Physical functioning	522 (74.9%)	157 (22.5%)	9 (1.3%)	9 (1.3%)	697
Role limitations	427 (61.3%)	221 (31.7%)	45 (6.5%)	4 (0.6%)	697
Social functioning	513 (73.6%)	167 (24.0%)	14 (2.0%)	3 (0.4%)	697
Pain	575 (82.5%)	101 (14.5%)	15 (2.2%)	6 (0.9%)	697
Mental health	398 (57.1%)	260 (37.3%)	36 (5.2%)	3 (0.4%)	697
Vitality	490 (70.3%)	191 (27.4%)	15 (2.2%)	1 (0.1%)	697

Table 4 presents different measures of agreement between the instruments for each dimension. With the exception of the mental health dimension, when comparing SF-6Dv2$_{\text{SF-36}}$ and SF-6Dv2$_{\text{ind-6}}$, Spearman's correlation coefficients were always higher than 0.7 and the mean weighted Kappa was 0.79 (range 0.61-0.91). Analysis of the results of the global consistency index (GCI), which computed the percentage of individuals classified in the same level of each dimension in the instruments compared, revealed a high level of agreement in responses with a mean of 70% (range 54.1-83.1). The best agreement in each level was found in the categories describing the best level (i.e., level 1 – no problem), while the ICI results decreased for the most severe levels (i.e., levels 5 and 6). Results suggested that even at these levels the degree of agreement was good, with the lowest degree found for the mental health dimension and the highest for the pain dimension.

Table 4. Rank correlations, Kappa, GCI and ICI between SF-6Dv2$_{\text{SF-36}}$, SF-6Dv2$_{\text{ind-10}}$ and SF-6Dv2$_{\text{ind-6}}$

Dimensions	**Spearman's correlation**	**Weighted Kappa**	**GCI**	**ICI (n)**					
				1	2	3	4	5	6
SF-6Dv2$_{\text{SF-36}}$ and SF-6Dv2$_{\text{ind-10}}$									
Physical functioning	0.858*	0.823	80.63	92.6 (215)	83.3 (287)	67.1 (143)	56.4 (39)	46.2 (13)	-
Role limitations	0.754*	0.762	60.26	63.3 (215)	60.2 (161)	61.5 (218)	52.6 (76)	40.7 (27)	-
Social functioning	0.762*	0.839	75.61	85.9 (327)	65.5 (177)	68.1 (138)	66.7 (39)	62.5 (16)	-
Pain	0.915*	0.911	82.07	85.7 (105)	80.6 (201)	76.5 (149)	87.1 (155)	85.3 (68)	68.4 (19)
Mental health	0.762*	0.756	66.71	76.7 (116)	67.5 (234)	66.5 (248)	58.5 (82)	29.4 (17)	-
Vitality	0.821*	0.799	68.58	75.3 (154)	57.3 (211)	74.9 (231)	67.1 (76)	68.0 (25)	-
SF-6Dv2$_{\text{SF-36}}$ and SF-6Dv2$_{\text{ind-6}}$									
Physical functioning	0.761*	0.728	71.45	85.6 (215)	85.0 (287)	30.0 (143)	59.0 (39)	30.8 (13)	-
Role limitations	0.706*	0.720	58.39	72.1 (215)	53.4 (161)	51.4 (218)	56.6 (76)	40.7 (27)	-
Social functioning	0.826*	0.826	74.46	89.0 (327)	58.8 (177)	66.7 (138)	66.7 (39)	37.5 (16)	-
Pain	0.910*	0.904	83.07	89.5 (105)	81.1 (201)	80.5 (149)	82.6 (155)	88.2 (68)	73.7 (19)
Mental health	0.664*	0.607	54.09	86.2 (116)	53.0 (234)	48.8 (248)	36.6 (82)	11.8 (17)	-
Vitality	0.749*	0.752	65.42	66.9 (154)	56.4 (211)	73.2 (231)	69.7 (76)	48.0 (25)	-
SF-6Dv2$_{\text{ind-10}}$ and SF-6Dv2$_{\text{ind-6}}$									
Physical functioning	0.775*	0.755	74.89	82.8 (233)	88.8 (285)	36.0 (125)	69.2 (39)	26.7 (15)	-
Role limitations	0.734*	0.732	61.26	81.4 (172)	55.0 (211)	55.5 (209)	55.6 (81)	41.7 (24)	-
Social functioning	0.818*	0.821	73.60	89.1 (321)	55.9 (186)	65.4 (136)	69.0 (42)	41.7 (12)	-
Pain	0.894*	0.887	82.50	91.7 (96)	80.8 (193)	80.3 (157)	80.0 (165)	86.9 (69)	76.5 (17)
Mental health	0.747*	0.696	57.10	92.0 (125)	53.2 (233)	48.8 (242)	46.4 (84)	15.4 (13)	-
Vitality	0.818*	0.830	70.30	73.2 (142)	63.8 (196)	76.1 (251)	76.7 (73)	56.0 (25)	-

* P < 0.01

In Table 5 we found that the indexes (i.e., the utility value for each instrument) provided very similar values. There was no significant difference between SF-6Dv2$_{\text{SF-36}}$ and SF-6Dv2$_{\text{ind-10}}$ (mean difference ± SD: -0.002 ± 0.070, p=0.38), but significant difference were observed for SF-6Dv2$_{\text{SF-}}$

$_{36}$ and SF-6Dv2$_{ind-6}$ (0.013 ± 0.075, p<0.01) and between SF-6Dv2$_{ind-10}$ and SF-6Dv2$_{ind-6}$ (0.016 ± 0.077, p<0.01). Similar results can be observed when differences are presented by sociodemographic characteristics (see Table A4 in appendix). Table 5 also indicates negligible and very similar floor (near 0) and ceiling effects (around 2%) in each index. Finally, Pearson's and intra-class correlation coefficients showed good to very good correlations (range 0.73-0.85). These results were also confirmed by the graphical plots of the three indexes provided in appendix where observations are well distributed along the diagonal.

Table 5. Descriptive statistics of the three indexes using the value set developed in the United Kingdom

	SF-6Dv2$_{SF-36}$	SF-6Dv2$_{ind-10}$	SF-6Dv2$_{ind-6}$
Observed range (theoretical: 0.29 - 1.00)	0.29 - 1.00	0.31 - 1.00	0.31 - 1.00
Mean (SD)	0.731 (0.128)	0.729 (0.130)	0.745 (0.130)
Median (IQR)	0.732 (0.64 - 0.82)	0.724 (0.64 - 0.82)	0.737 (0.66 - 0.84)
Kurtosis	0.0239	0.0851	0.0016
Skewness	0.3728	0.632	0.183
Ceiling effect (%)	2.01	1.87	2.58
Floor effect (%)	0.14	0.00	0.00
Mean difference with SF-6Dv2$_{SF-36}$	-	-0.002	0.013*
Mean difference with SF-6Dv2$_{ind-10}$	0.002	-	0.016*
Median difference with SF-6Dv2$_{SF-36}$	-	-0.008	0.005
Median difference with SF-6Dv2$_{ind-10}$	0.008	-	0.013*
Pearson's correlation with SF-6Dv2$_{SF-36}$	-	0.8515*	0.8299*
Pearson's correlation with SF-6Dv2$_{ind-10}$	0.8515*	-	0.8232*
Intraclass correlation coefficient SF-6Dv2$_{SF-36}$	-	0.733*β	0.776*
Intraclass correlation coefficient SF-6Dv2$_{ind-10}$	0.733*β	-	0.827*

* P < 0.01; β ICC was calculated for the three indexes together.

It may be noted that, in very uncommon cases, the intra-class correlation coefficient can become negative when the within-groups variance exceeds the between-groups variance. In such cases, the ICC is not appropriate and it is better to consider the Pearson's correlation coefficient. However, to overcome this problem it is also possible to consider the ICC for the three groups as a proxy.

Two additional analyses were performed to evaluate if the order of administration had an impact on results: 1) SF-6Dv2$_{SF-36}$ systematically administered first and SF-6Dv2$_{ind-10}$ second (n=112); and 2) SF-6Dv2$_{SF-36}$ systematically administered first and SF-6Dv2$_{ind-6}$ second (n=114). In these two subgroups, the degree of agreement remained very good for weighted Kappa, Spearman's correlation coefficients and GCI. However, the ICI performed poorly than in the whole sample, particularly in the second subgroup (see Table A5 in appendix). As regards to the utility values generated in these two sub-groups, higher values were generated with SF-6Dv2$_{ind-10}$ and particularly with SF-6Dv2$_{ind-6}$ (see Table A6 in appendix).

Discussion

In this study, we tested if the different formats in which the SF-6Dv2 can be used provide consistent results and if they can be used interchangeably. Exploring the consistency of the SF-6Dv2, we found very similar results for the three instruments compared, with high degrees of agreement, high levels of correlation, and low mean difference. In addition, we found very little ceiling effect among the three instruments.

The same methodology as Ferreira et al. [10] was used with the differences that: 1) the instruments were administered in a random order; 2) we used the new version of the SF-6D (i.e., SF-6Dv2); 3) we also tested if the 10 items from the SF-36v2 can be used alone to calculate a utility value or if it is necessary to complete the full questionnaire with 36 items (i.e., not done in the study by Ferreira et al. [10]); and 4) an additional agreement measure was used with the quadratic weighted Kappa.

Our results were considerably better than those found in the study by Ferreira et al. [10]. With the exception of the mental health dimension, Spearman's correlation coefficients were always higher than 0.7, which was never the case in Ferreira et al. [10]. In addition, the GCI percentages are very high as compared to those found in Ferreira et al. [10] with a mean difference of 23 percentage points, which is more than double for the dimensions of mental health and vitality. This evidence of strong agreement between responses was also supported by the results of the ICI, which defines the level of stability in responses and is calculated as the percentages of individuals correctly classified in a particular level of each dimension. An explanation for this difference can be found in the fact that the three instruments were administered in a random order in our study, while the SF-6Dv1$_{\text{SF-36}}$ was systematically administered first in the study by Ferreira et al. [10]. This could have introduced a systematic bias towards the use of the SF-6Dv1 with the full SF-36v2 questionnaire.

When exploring the impact of order of administration, the resuls of comparisons in SF-6Dv2$_{\text{SF-36}}$ systematically administered first and SF-6Dv2$_{\text{ind-6}}$ second matched the most with the poor results reported by Ferreira et al. [10]. However, the degree of agreement still remained very good for Spearman's correlation coefficients, weighted Kappa and GCI. Only the ICI performed much more poorly than in the whole sample and higher values were generated for utility values. Although these differences in utility values were about twice more important than those found in complete data, these differences were well below the results of the study by Ferreira et al. [10]. In light of these elements, it seems that the order of administration of the instrument could only partially explain the results reported by Ferreira et al. [10]. Consequently, there may be other reasons. For example, our sample was much more representative of the general population and the number of observations was higher, which could have improved our results. Specifically, Ferreira et al. [10] found significant differences between the utility values across sociodemographic groups, which was not the case in our study. Also, one should consider that we used a different population than Ferreira in terms of region and culture, which may have impacted our results in an unpredictable way. However, considering that our sample population was older and more representative of the general population than the sample collected by Ferreira et al. [10], our study pertained to a higher proportion of severe health states, but also to a better distribution of ratings. This may explain why we found no or little ceiling effect in our sample. These elements and the fact that the degree of agreement between the instruments was higher in our study reinforce our confidence in our results. Indeed, with a better distribution of ratings, we had a higher probability that a respondent changed a rating, especially in the middle of the scale, but this was not the case. Finally, in this study we used the SF-6Dv2, an improved version of the SF-6D with better psychometric characteristics [9], which may have influenced our results in a positive way.

Considering the SF-6Dv2$_{\text{SF-36}}$ as the reference, the worst corresponding results in our study were obtained for the SF-6Dv2$_{\text{ind-6}}$, but they were still very good with very high correlation and agreement with the SF-6Dv2$_{\text{SF-36}}$ and a small mean difference in utility values. In addition, the highest significant difference was found between the SF-6Dv2$_{\text{ind-10}}$ and SF-6Dv2$_{\text{ind-6}}$ with a mean difference of 0.016. Even if statistically significant ($p<0.01$), this difference is very small and well below the mean minimally important difference (MID) of 0.041 reported by Walters and Brazier [13] for a change in QALY. As compared to the results of Ferreira et al. [10], our results indicated a difference 4 to 60 times less important. When comparing the index SF-6Dv2$_{\text{SF-36}}$ with SF-6Dv2$_{\text{ind-6}}$, as Ferreira et al. [10] did, the difference in our sample was only of 0.013, which is 9 times less important with the same value set used to convert into QALY (i.e., UK value set). We believe that the minor differences between instruments found in our study were mainly due to random errors with respondents simply making mistakes when answering each question. These results thus indicate that the SF-6Dv2 may potentially be used as an independent instrument with minimal bias, both for SF-6Dv2$_{\text{ind-10}}$ and SF-6Dv2$_{\text{ind-6}}$. Indeed, our results indicate that to complete the classification system of the SF-6Dv2 alone will lead to small differences in responses and utility values.

In addition, since the utilities from the UK value set have been generated using direct elicitation techniques with the phrasing of the 6 dimensions of the SF-6Dv2 and not with the phrasing of the items in the SF-36v2, it may be deemed paradoxical that the SF-6D should be administered through the SF-36v2. Historically, this was justified by the fact that the SF-6D was designed to derive utilities from the SF-36v2. However, if we can use the SF-6Dv2 as an independent instrument without changing the results in utility values, this can change a lot in how surveys are conducted; particularly, it will considerably reduce the length of the questionnaire (i.e., 6 questions instead of 36) and fatigue to patients. However, we should remain cautious since the mean differences values obtained in this study can hide small but significant discrepancies between respondents' answers, as shown in Table 4.

A limit in our study is that we do not know if we would get the same results in another context (i.e., outside Quebec), since how to complete the SF-6Dv2 instruments could be different among different people because of the way they interpret the questions and possible answers. One study is not sufficient and others should be conducted to confirm these results. Moreover, it will be necessary to redo this analysis when a value set is available with the SF-6Dv2, and particularly for Quebec. Considering that health preferences can be different between countries [8], it is necessary to have a value set for Quebec to confirm our results. Furthermore, some authors indicate that it would also be necessary to assess the preferences of sick populations [14-15], but actually it will in contradiction with the theoretical model of QALY which aims to help a decision-making based on the point of view of the general population (i.e. the tax-payer) [3]. Another limit of our study is that we did not compare the SF-6Dv1 with the SF-6Dv2, which may have contributed to assess the validity of the new SF-6D version.

To conclude, even if the use of the full SF-36v2 questionnaire is still recommended to generate utility values from the SF-6D, this study provides evidence that the SF-6Dv2 may be used as an independent instrument with minimal bias, either in combination with the 10 items from the SF-36v2, or with the classification system used in the SF-6Dv2 (i.e., 6 questions).

References

[1] Weinstein MC, Torrance G, McGuire A. QALYs: The Basics. Value in Health 2009;12(S1):S5-S9.

[2] Richardson J, McKie J, Bariola E. Multiattribute Utility Instruments and Their Use. In Encyclopedia of Health Economics, vol. 2. Editors: Anthony J. Culyer. San Diego: Elsevier; 2014:341-357.

[3] Fauteux V, Poder TG. État des lieux sur les méthodes d'élicitation du QALY [Overview of QALY elicitation methods]. Int J Health Pref Research 2017;1 :2-14.

[4] Ungar J. Challenges in Health State Valuation in Paediatric Economic Evaluation. PharmacoEconomics 2011:29(8):641-652.

[5] Brazier JE. The Short-Form 36 (SF-36) Health Survey and Its Use in Pharmacoeconomic Evaluation. PharmacoEconomics 1995;7(5):403-415.

[6] Brazier JE, Usherwood T, Harper R, Thomas K. Deriving a Preference-Based Single Index from the UK SF-36 Health Survey. J Clin Epidemiol 1998;51(11):1115–1128.

[7] Brazier JE, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 2002;21:271-292.

[8] Poder TG, Gandji EW. SF-6D value sets: a systematic review. Value in Health 2016;19(3):A282.

[9] Mulhern B, Brazier J. Developing version 2 of the SF-6D: The health state classification system. In: 21st Annual Conference of the International Society for Quality of Life Research. Qual Life Res; 2014; 23(S1):49.

[10] Ferreira LN, Ferreira PL, Pereira LN, Rowen D, Brazier JE. Exploring the consistency of the SF-6D. Value in Health 2013;16(6):1023-1031.

[11] Bansback N, Mulhern B, Sawatsky R, Brazier JE, Whitehurst D. Valuing the SF-6Dv2 in Canada. Quality of Life Research 2015;24(1):180-181.

[12] Pickard AS, Wang Z, Walton SM, Lee TA. Are decisions using cost-utility analyses robust to choice of SF-36/SF-12 preference-based algorithm? Health and Quality of Life Outcomes 2005;3:11.

[13] Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005;14:1523–1532.

[14] Versteegh MM, Brouwer WBF. Patient and general public preferences for health states: A call to reconsider current guidelines. Soc Sci Med. 2016;165:66–74.

[15] De Pouvourville G. Editorial: Whose preferences count in decision-making? Int. J. Health Pref. Researc. 2018, 1:1-2