ANALYSIS OF EXPERT RATINGS ON HEALTH-RELATED QUALITY OF LIFE METRICS

Bryce Mason, Ph.D.[1]
RAND
May 2005

This paper reports the results of statistical analyses of the expert assignments of attributes under the health-related quality of life (HRQL) indexes used in the NHTSA case study (Robinson, Corso et al., 2005). In particular, it assesses expert agreement *within* and *across* indexes; that is, how well experts agreed on HRQL attributes on a given index (i.e., a reliability assessment) and how well experts agreed on overall quality-adjusted life year (QALY) losses, irrespective of the index used. This document also describes the statistical analyses used and how the available data affected the types of analyses possible.

Data were collected for four HRQL indexes, but the data were not consistently collected across index. The EQ-5D and HUI-2 had the most complete data, with four experts rating 20 unique cases; 16 were classified as nonfatal and 4 were classified as fatal. One unrated fatal case and one duplicate injury case were excluded (cases 15 and 17 had the same injury description; see Exhibit 4 in the case study). Although five cases were fatal, experts assigned attributes to four cases as if the child survived the accident.[2] Nevertheless, each of the four experts provided overall assessments of each case's health in the acute, rehabilitation, and long-term phases. With respect to the QWB, HRQL assessments were made on the same 20 cases, but QWB index values had been estimated in prior research for adult patients and aggregated by body part affected and Abbreviated Injury Scale (AIS) score (Holbrook, et al., 1999); the injury with the lowest QWB value was used in this statistical analysis.[3] Thus case-level HRQL in the long-term phase was imputed as being equal to the minimum HRQL value of the injuries for a given case. Case-level, long-term (i.e., 12 months post injury) FCI estimates were provided by MacKenzie for the 16 non-fatal cases and were based on AIS descriptions for each injury (see MacKenzie, et al., 1996). As discussed in Robinson, Corso et al. (2005), experts assigned injuries described by the AIS to the domains and attribute levels of the FCI survey instrument.

The different available data had important ramifications on the type of statistical analyses that could be used to assess expert agreement. In particular, the highest-quality judgments about within-measure reliability could only be made about the HUI-2 and the EQ-5D, since the same experts provided independent ratings on each measure, as opposed to an average rating from a different set of experts. However, because estimates of long-term HRQL existed or could be constructed from each of the indexes, some judgments about the long-term agreement in HRQL across indexes could also be made.

[1] Author email: brycemason@gmail.com
[2] Under the Abbreviated Injury Scale (AIS), the case number 5 was placed in Category 6; i.e., as generally leading to immediate fatality, and was not included in the expert assignment process.
[3] The case study uses the injury as posing the largest fatality risk (i.e., the maximum AIS), which is not always the same as the injury with the larges HRQL impacts.

In general, experts assigned attributes under the HUI-2 and EQ-5D to each of the 20 cases for the acute, rehabilitation, and long-term phases of each injury. These attributes were then valued based on the standard community preference surveys associated with each index, and the results were multiplied by duration (taking life expectancy into account) and aggregated across the different injury phases for each case. So, for these indexes, there were two types of ratings available for each phase and each injury case: attribute assignments reflecting descriptive information on HRQL impacts, and weighted index values reflecting the relative value of these impacts and their duration. These two types of data warranted different statistical approaches to establish agreement.

First, the kappa-statistic (Landis and Koch, 1977) of interrater agreement was used to calculate the agreement between experts within a particular phase, dimension, and HRQL index; i.e., agreement on the attributes assigned to each time period under each index for each injury case. This statistic is suited to categorical data—where a fixed number of experts have classified observations into a fixed number of categories. The statistic is scaled to be zero when the agreement between experts is what is expected by chance, and unity when there is perfect agreement. Importantly, if raters all chose a particular value with great frequency and there is little variation in observations, kappa tends to be low since the a priori probabilities of each rating are calculated using the frequency in the data. In other words, if all experts rated a particular case as a "1" in some dimension then kappa would be "0" since there was no variation—kappa can underestimate agreement when there is little variation in ratings.

Landis and Koch (1977) suggest the following interpretations of kappa:

Below 0.0	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.80 – 1.00	Almost perfect

Once the HRQL formulae were applied to the individual attributes and expressed as overall HRQL values, kappa was no longer applicable because there were no clear categories. HRQL and QALY values are pseudo-continuous measures for which simple correlations (rho) were calculated between measures. The table above is still useful in interpreting rho.

Table 1 presents basic summary statistics for the expert rating data: mean values, standard deviations and number of observations for all the variables. For each variable, the total possible number of observations is 80 (20 cases times four experts), although there were a few cases of missing data, especially in the fertility domain of the HUI-2. For these analyses the simple imputation rule applied elsewhere in the case study was used; every missing rating was set to no impairment, or a value of 1. One expert had zero duration for the rehabilitation phase for several cases, and did not fill out the cells for those cases. In the one case of missing duration data, the duration from the other index was used (same expert, same phase). It is important to note that experts were not always consistent in their estimates of phase duration across the two generic indexes, despite the fact that the injuries were identical.

Table 1. Descriptive Statistics of Experts' Ratings, by Index and Phase						
	Acute Phase		**Rehabilitation Phase**		**Long Term Phase**	
	n	**mean (SD)**	**n**	**mean (SD)**	**n**	**mean (SD)**
HUI-2						
Duration (years)	80	0.06	80	0.19	N/A	N/A
		(0.05)		(0.23)		
Sensation	80	1.95	77	1.61	80	1.32
		(1.18)		(0.92)		(0.69)
Mobility	80	3.28	77	2.40	79	1.53
		(1.67)		(1.28)		(1.06)
Emotion	80	2.49	77	1.87	79	1.47
		(1.30)		(0.92)		(0.78)
Cognitive	80	2.29	76	1.88	79	1.52
		(1.24)		(0.99)		(0.77)
Self-care	80	3.01	76	2.41	78	1.54
		(1.17)		(1.15)		(0.98)
Pain	76	3.41	77	2.13	79	1.37
		(1.00)		(0.71)		(0.62)
Fertility	59	1.41	77	1.40	79	1.21
		(0.75)		(0.69)		(0.50)
EQ-5D						
Duration (years)	80	0.06	79	0.18	N/A	N/A
		(0.05)		(0.23)		
Mobility	80	2.13	77	1.71	80	1.26
		(0.81)		(0.69)		(0.55)
Self-care	79	2.23	77	1.87	80	1.31
		(0.77)		(0.75)		(0.63)
Usual activities	79	2.41	77	2.05	80	1.43
		(0.69)		(0.69)		(0.69)
Pain/Discomfort	79	2.27	77	1.91	79	1.18
		(0.61)		(0.46)		(0.38)
Anxiety/Depression	79	1.81	77	1.60	79	1.38
		(0.70)		(0.61)		(0.63)

The interrater agreement between every expert for each index and phase is presented in Tables 2 and 3. In these statistics, disagreements were weighted by how “close” they were. For example, if two experts rated mobility in EQ-5D for a certain case and phase (with possible rating 1, 2, or 3), they got 1 point for exact agreement, 1/2 point if their ratings differed by 1, and 0 points if one rated a 1 and the other a 3. Because the kappa statistic shows to what degree actual agreement is better than expected random agreement, this method does not inflate values. For the HUI-2, mobility, self-care, and sensation were the three domains with the most agreement. In the matrices describing these relationships, kappa is almost always greater than 0.50. Some domains had less agreement, such as pain and emotion (in the acute and rehabilitation phases), although experts tended to agree on both in the long-term phase (emotion in the long term has low kappa, but this is an artifact of the statistic—virtually everyone marked no impact of the accident on emotion in the long run).

Table 2. Complete HUI-2 Interrater Analysis

Sensation	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.66				0.84				0.45			
C	0.47	0.64			0.44	0.34			0.44	0.33		
D	0.64	0.82	0.54		0.78	0.85	0.49		0.59	0.31	0.53	
Mobility	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.73				0.46				0.53			
C	0.59	0.61			0.35	0.38			0.63	0.71		
D	0.68	0.65	0.47		0.51	0.56	0.33		0.60	0.43	0.43	
Emotion	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.23				0.33				0.45			
C	0.19	0.42			0.41	0.35			0.46	0.30		
D	0.00	0.26	0.16		0.00	0.33	0.05		0.55	0.77	0.49	
Cognitive	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.82				0.84				0.63			
C	0.73	0.76			0.64	0.62			0.71	0.46		
D	0.51	0.44	0.44		0.27	0.40	0.13		0.79	0.67	0.76	
Self-care	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.74				0.35				0.66			
C	0.62	0.69			0.29	0.54			0.63	0.52		
D	0.67	0.48	0.46		0.67	0.58	0.39		0.56	0.51	0.46	
Pain	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.05				0.12				0.00			
C	0.06	0.00			0.41	0.18			0.00	0.09		
D	0.21	0.00	0.09		-0.02	-0.04	0.30		0.00	0.20	0.04	
Fertility	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.00				0.00				0.00			
C	0.00	0.63			0.00	0.34			0.00	0.36		
D	0.00	0.00	0.00		0.00	0.63	0.50		0.00	0.46	0.83	

Note: All values greater than 0.10 were significant at the 5% level.

Table 3. Complete EQ-5D Interrater Analysis												
Mobility	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.67				0.46				0.62			
C	0.89	0.78			0.64	0.60			0.27	0.14		
D	0.45	0.36	0.48		0.32	0.49	0.69		0.67	0.41	0.24	
Self-care	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.52				0.58				0.71			
C	0.70	0.71			0.15	0.41			0.16	0.14		
D	0.36	0.15	0.33		0.57	0.39	0.29		0.59	0.73	0.23	
Usual activities	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.51				0.67				0.73			
C	0.59	0.67			0.21	0.29			0.33	0.29		
D	0.71	0.38	0.31		0.30	0.33	0.32		0.86	0.57	0.41	
Pain	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.21				0.41				0.18			
C	-0.08	0.53			0.77	0.31			0.00	0.00		
D	0.31	-0.06	-0.08		0.31	0.29	0.26		-0.09	0.43	0.00	
Anxiety / depression	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.65				0.71				0.66			
C	0.33	0.30			0.32	0.42			0.00	0.00		
D	0.46	0.38	0.20		0.30	0.45	0.64		0.56	0.73	0.00	

Note: All values greater than 0.10 were significant at the 5% level.

These matrices are probably too detailed for all but the most interested reader, so Table 4 summarizes the information in Tables 2 and 3 two ways. Table 4 contains the interrater agreement within each domain and across all the phases (so there are 20 x 3 cases, each rated by 4 raters). The table shows somewhat less agreement for emotion, pain and fertility (.16 - .19) but about the same moderate level of agreement for the other nine dimensions (.37 - .52). It appears that the assignment of EQ-5D attributes reflects more agreement than the HUI-2 on average.

Table 4. Summary of Expert Agreement (Interrater Agreement Across all Phases—kappa)	
HUI-2:	
Sensation	0.45
Mobility	0.39
Emotion	0.19
Cognitive	0.41
Self-care	0.46
Pain	0.16
Fertility	0.19
EQ-5D:	
Mobility	0.52
Self-care	0.42
Usual activities	0.46
Pain / Discomfort	0.40
Anxiety / Depression	0.37

Table 5 summarizes the information a different way that might be more relevant to the calculation of QALYs. This time, HRQL summary valuations combining the various domains were correlated across all experts and a matrix was used to display these relationships. These values were not multiplied by duration (i.e., they are not QALY estimates). Here we can see a high degree of correlation overall, although some pairs of experts tended to have overall HRQL valuations of health that were more similar than others. For example, A's and B's estimates resulted in HRQL values that were correlated higher than the median rho of .77 in all but one case, whereas C and D had less correlated HRQL estimates. There tended to be more agreement in total HRQL valuation in the long-term phase than all other phases, and the overall correlations were quite high.

Table 5. Interrater Correlations of HRQL Index Values by Phase and Measure

HUI-2	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.90				0.87				0.77			
C	0.57	0.77			0.70	0.74			0.69	0.72		
D	0.76	0.69	0.53		0.88	0.87	0.78		0.83	0.83	0.76	
EQ-5D	*Acute*				*Rehabilitation*				*Permanent*			
	A	B	C	D	A	B	C	D	A	B	C	D
A												
B	0.67				0.79				0.96			
C	0.82	0.85			0.77	0.80			0.92	0.83		
D	0.59	0.32	0.38		0.50	0.67	0.72		0.95	0.90	0.96	

To address the question of whether there were systematic differences between the EQ-5D and HUI-2 estimates, estimates of HRQL and QALYs lost (in the first two time periods) were correlated across indexes. This analysis had 80 observations for each phase (20 cases, four experts), but only the acute and rehabilitation phase were considered in Tables 6 through 8. First, Table 6 summarizes the correlation between the two indexes. The two indexes lead to highly correlated valuations and estimates of QALYs lost across the two phases. The QALYs lost as estimated by each index are more highly correlated than the valuations because of the very high correlation in the duration estimated for each HRQL index. The most important correlation in this analysis—QALYs lost across the acute and rehabilitation phases—is almost perfectly correlated at .92. Thus, the two indexes order the cases in almost exactly the same way. Correlation is a good measure of whether experts order the cases in the same way, but does not show whether one expert had systematically higher or lower estimates of lost health per case; one might want to know how much the experts disagree from each other.

Table 6. Correlation between HUI2 and EQ-5D, by Phase

	Phase		
	Acute	*Rehabilitation*	*Acute + Rehabilitation*
Duration (D)	0.99	0.99	0.99
HRQL (V)	0.70	0.83	
QALYs Lost (D * (1-V))	0.93	0.91	0.92

Notes:
All correlations were significant at the 1% level. N = 80 per phase.
This calculation compares the “with condition” HRQL to perfect health (a value of 1.0).

These systematic differences are examined in the ANOVA regressions in Table 7 and 8. The regression in Table 7 relates duration estimates with HRQL index and expert. The differences in the duration estimates for the two indexes were negligible (the experts gave an extra .59 days on average when filling out the HUI-2 form—indeed, it might have been better just to ask for duration once, insuring that there would be no difference). However, there were substantial differences across experts’ estimates of duration. Using expert A as the baseline, the duration estimates from the other experts averaged between two and 23 days less in the acute phase. The differences between experts’ ratings of duration were even larger in the rehabilitation phase, ranging from 19 days less than A to 75 days more. This might not be a surprise since the period of rehabilitation is in general longer and more variable than the acute phase (as seen in the constant term, reflecting expert A’s estimate). Adding coefficients across the two phases, we see total estimated duration differences were 26, 69, and -43 days for B, C, and D respectively. These appear to be large differences.

Table 7. Duration Estimation Differences as a Function of HRQL Index and Expert			
Baseline = EQ5D, Expert A			
Acute Phase	**Coef.**	**t**	**p**
HUI2	0.59	0.25	0.81
Expert			
B	-2.61	-0.78	0.44
C	-6.70	-1.99	0.05
D	-23.29	-6.91	< 0.01
Constant	30.02	11.27	< 0.01
Rehabilitation Phase	**Coef.**	**t**	**p**
HUI2	1.61	0.13	0.90
Expert			
B	28.92	1.67	0.10
C	75.77	4.38	< 0.01
D	-19.37	-1.12	0.27
Constant	45.67	3.34	< 0.01
Note: Duration was measured in days.			

The regression in Table 8 estimates the difference between quality-adjusted life days (QALDs = QALYs / 365.25) lost in the combined acute and rehabilitation phase, according to the index and each of the experts. The difference between the two indexes is, on average, only about two QALDs, and this difference was statistically insignificant. However, there is clear variation between experts as indicated by the coefficients on the expert indicator variables. The differences are highly related to aforementioned differences in the duration for the two phases. None of them attained significance, but this is almost assuredly due to low power.

Table 8. QALYs Lost as a Function of HRQL Index and Expert			
Baseline = EQ5D, Expert A			
	Coef.	**t**	**p**
HUI2	-2.05	-0.20	0.84
Expert			
B	6.93	0.48	0.63
C	19.31	1.33	0.19
D	-24.65	-1.70	0.09
Constant	48.68	4.25	< 0.01
Note: Outcomes are in quality adjusted life days (QALDs = QALY * 365.25)			

Finally, estimates of long-term health outcomes were compared according to all four HRQL indexes, considering the QWB and FCI in addition to the EQ-5D and HUI-2. To answer this question, HRQL for the long-term phase was estimated and correlated across the four measures. As mentioned earlier, the QWB and FCI index values were based on different estimation methods, reducing the validity of the overall comparison. In order to make the indexes comparable, values for the long-term phase were averaged across the four experts for the 20 cases included in the HUI-2 and EQ-5D analysis.[4] The sources of the QWB and FCI values differ significantly as discussed in the case study (Robinson, Corso et al., 2005). For the FCI, 12-month values (as reported in Exhibit 14 of the case study) for the five cases expected to impact long-term HRQL were used; for the remaining nonfatal cases, an HRQL value of 1.0 (perfect health) was used. Again, one duplicate nonfatal case and all fatal cases were dropped (FCI values were not available for fatal cases). For the QWB, the 18-month values for the worst injury in each of the 20 cases included in the EQ-5D and HUI-2 analysis were used, based on the mean adult estimates (reported in Exhibit 13), again treating the four fatal cases as if they were nonfatal. Correlations were estimated despite these differences in how long-term HRQL was calculated. This had to be done twice, once on the cases included in the FCI estimates (which reduced the sample size to 16 cases per index), and once for the other three measures on the complete sample (20 cases per index). Table 9 shows the difference in long-run HRQL according to EQ-5D, HUI-2, FCI (at 1 year), and QWB (at 18 months). Once again, the EQ-5D and HUI-2 were very highly correlated and the other measures were more moderately correlated.

Table 9. Correlation of Long-term HRQL across EQ-5D, HUI2, FCI, and QWB Indexes*

Without FCI (20 cases):	*EQ-5D*	*HUI-2*	*QWB*	
EQ-5D				
HUI-2	*0.97			
QWB	*0.48	*0.45		
With FCI (16 cases):	*EQ-5D*	*HUI2*	*FCI*	*QWB*
EQ-5D				
HUI-2	*0.98			
FCI	*0.90	*0.92		
QWB	0.39	0.39	0.42	

Notes:
* Correlation significant at the 5% level.
See Robinson, Corso et al. (2005) for discussion of sources of estimates for each index.
Correlations calculated with and without the FCI measure as FCI estimates are available for only 16 cases, reducing the analysis to 16 observations per index when the FCI was included.

Table 10 shows the mean differences in long-term HRQL according to the four indexes as a result of an ANOVA regression (again the ANOVA was conducted twice, once with the FCI reduced sample and once without the FCI on the complete sample). In the long-run phase, the average HRQL was high (>.85) and there was no statistical difference between any measure except the QWB, which had systematically lower scores. This table suggests no practical difference in long-term HRQL between the EQ-5D, HUI-2, and the FCI.

[4] The case study (Robinson, Corso et al., 2005) relies on median estimates.

Table 10. HRQL differences as a function of HRQL Index			
Without FCI (Baseline = EQ-5D, n = 20 cases per index)			
	Coef.	t	p
HUI2	0.01	0.19	0.85
QWB	-0.17	-2.93	< 0.01
Constant	0.84	20.61	< 0.01
With FCI (Baseline = EQ-5D, n = 16 cases per index)			
	Coef.	t	p
HUI2	0.02	0.37	0.71
FCI	0.04	0.69	0.49
QWB	-0.20	-3.91	< 0.01
Constant	0.88	23.89	< 0.01
Note: Coefficients show mean differences between EQ-5D and other HRQL indexes.			

Although there were differences between expert ratings of the HRQL attributes for these crash-related injuries along the domains of the EQ-5D and HUI-2, there was almost no difference in the estimates of QALYs lost in the acute and rehabilitation phases across the indexes. However, there were differences among experts. In other words, there was substantial variation between experts, presumably due to how bad they thought the consequences of the injuries would be, but it did not matter whether they used the EQ-5D or HUI-2 to express those consequences. The major difference in the estimates of QALYs lost between the experts was due to varying estimates of the duration of acute and rehabilitation phases, rather than in their estimates of HRQL in those states. The HRQL differences might have been reduced further by more training or agreements on how, for example, the experts were supposed to rate the fertility effects on these young children during the acute phase of the injury.

References

Holbrook, Troy L., John P. Anderson, William L. Seiber, Dierdre Browner, and David B. Hoyt. 1999. Outcome After Major Trauma: 12 Month and 18 Month Follow-Up Results from the Trauma Recovery Project. *The Journal of Trauma: Injury, Infection and Critical Care.* Vol. 46: 5, pp. 765-773.

Landis, J. R., and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:1, pp. 159-174

MacKenzie, Ellen J., Anne Damiano, Ted Miller, and Steve Luchter.1996. The Development of the Functional Capacity Index. *The Journal of Trauma: Injury, Infection, and Critical Care.* Vol. 41: 5.

Robinson, Lisa A., Phaedra Corso, Xiangming Fang, Robert Black et al. 2005. Alternative Approaches for Estimating Health-Related Quality of Life Impacts: Child Restraints Regulation Case Study. Washington, DC: Institute of Medicine. Available at: www.iom.edu/file.asp?id=31437.