University of York Department of Health Sciences

Measurement in Health and Disease

Outcome Measures

David Torgerson
Director, York Trials Unit

(Written version by Martin Bland)

Surrogate measures

There are numerous methods of measuring outcomes in trials. Usually, we need to measure 'clinical' effects, such as blood pressure or survival, and quality of life. Often quality of life and clinical measures will be related, but they may not.

'Clinical' outcomes are numerous and are often 'surrogates' for 'real' outcomes. By this, we mean that the surrogate variable is a substitute for the variable which is really of interest. For example, in the study of vascular disease, the real variables of interest would be stroke, angina, heart attack, or death. We may actually measure blood pressure or blood lipids, because as quantitative variables we can find differences in much smaller samples and shorter follow-up times. In osteoporosis, we might be interested in preventing fractures but use surrogates such as bone mass and bone turnover. In a trial aimed at reducing partner assault we might want to achieve a reduction in assaults, but might measure changes in a questionnaire scale of attitudes to partnership. In evaluating an MSc lecture, we might want to know whether the lecture increased knowledge, but actually measure students' enjoyment and satisfaction instead.

There are problems with surrogates. First, change in surrogates may not lead to changes in real outcomes. For example, sodium fluoride increases bone mass but also increases fractures. Hormone replacement therapy (HRT) reduces cholesterol levels but increases risk of stroke and cardiac disease.

HRT profoundly affects a wide range of surrogates. It improves blood cholesterol, increases blood flow to the brain, and increase bone mass. However, trials with real outcomes shows increases in deaths due to cardiovascular disease and increased incidence of dementia. HRT does reduce fractures, but only one of the three surrogates correctly predicted an improvement in the real outcome.

The treatment of AIDS provides another example. Some successful anti-AIDS drugs have little or no effect on cellular markers of disease progression, but in trials of the drugs with AIDS death as the outcome they did reduce deaths.

Another often-used surrogate is user satisfaction. Some trials show either qualitatively or quantitatively an improvement in treatment satisfaction but no change in real outcome. For example, counselling for women after traumatic childbirth increases satisfaction with the service but also increases post natal depression.

This can work the other way round, too. A randomised controlled trial of the use of cognitive behavioural therapy (CBT) to increase the chance of finding a job showed no difference between the groups in 'job seeking activities' (e.g., number of interviews, number of job applications etc) but the trial showed those allocated to CBT were

significantly more likely to get work (34%) than the controls (13%) ($p < 0.001$). Had the trial only measured job seeking behaviour then we would have concluded, erroneously, that CBT was a useless intervention at increasing employment for the long term unemployed.

The Atkins Diet provides our final example. Dieticians dislike the Atkins Diet at it goes against 'accepted' wisdom. However, whilst weight loss isn't much different from a low carbohydrate diet lipids (surrogates) for cardiovascular disease are better.

It seems surrogates are mistrusted if they go against accepted wisdom but trusted if they confirm the prior hypothesis.

If surrogate markers can be misleading, why use them? Often cost is the explanation – real outcomes of death or disablement require huge expensive trials, whereas surrogate markers will tend to confirm that a drug is acting as theory suggests it should. For example, bone mass changes confirm the drug is reaching the bone and exerting an effect.

Another is the reliance on class effects. Often 'me to' drugs use markers as they act in a very similar way as established treatments and the assumption is made that they if they reduce the surrogate they will also reduce the real event. For example, daily bisphosphonate treatment increases bone mass and reduces fractures. Weekly treatment increases bone mass the assumption is that weekly treatment will reduce fractures as much as daily.

Sample size is also a factor. Usually surrogates need a much smaller sample size to show an effect, which reduces the cost, increases the speed of the trial etc.

Surrogate thus have several potential advantages, but we need to be wary of their use. We need to identify outcomes that are of interest to the patient, not the clinician, biologist or social scientist. Surrogate outcomes are rarely like this.

Quality of Life

The aim of most health care is to improve quality of life. For many people extending life or preventing death is not necessarily the most important aspect. Many treatments will extend life or increase the probability of survival but at the 'expense' of very poor quality of life. For example, radical surgery of head and neck cancer will improve survival from very low levels by only a small amount. The increase in survival time may carry with it terrible quality of life effects: patient can't speak properly, have difficulty eating, and may have terrible disfigurement. The majority of patients will still die quite soon but will have their remaining life span of very poor quality.

It is very important to be able to measure quality of life. A number of quality of life scales are widely used. We may classify them in three categories:

- Disease specific;
- Generic measures;
- Utility measures.

Figure 1. A section of SF36

Your Feelings					
(Please Circle **One Number on Each Line**)					
- These questions are about how you feel and how things have been with you **during the past month**. For each question please give the one answer that comes closest to the way you have been feeling. How much during **the last month**:	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Have you felt calm and peaceful?	1	2	3	4	5
b) Did you have a lot of energy?	1	2	3	4	5
c) Have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5

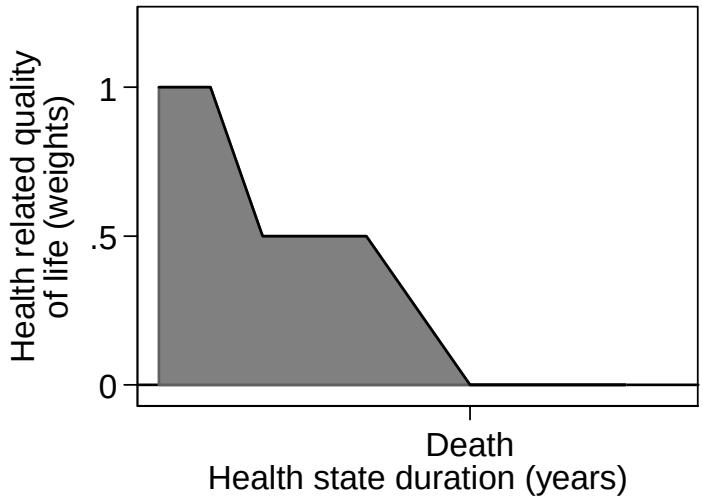
Disease specific measures are questionnaires that will ask specific questions relating to the health condition. For example, the Roland & Morris back pain scale asks 24 questions about disability related to your back (e.g., do you have trouble getting out of a chair because of your back pain?).

Disease specific measures have a number of advantages in that they are 'sensitive' to changes in the condition. However, they will not pick up other general health disadvantages or benefits of treatment. For example, they will not pick up cessation of depression through curing back pain.

Generic measures of health and quality of life have questions asking about general health (e.g., SF36, SF12, Nottingham health profile (NHP), Women's Health Questionnaire). The advantage is that they will pick up other effects of treatments. The disadvantage is that they may not be sensitive to small, but important, health effects related to the condition being studied.

One very widely used generic measure is the SF36 (or Short Form 36 item questionnaire, a name by which it is never known). Figure 1 shows a section of this 36 item questionnaire. The SF36 is one of the most widely used generic quality of life instruments. It is derived from the much longer Medical Outcome Survey Instrument developed by the RAND corporation as part of a massive RCT of payment systems for health care treatment.

Figure 2. Quality Adjusted Life Years (QALYs) for a hypothetical person



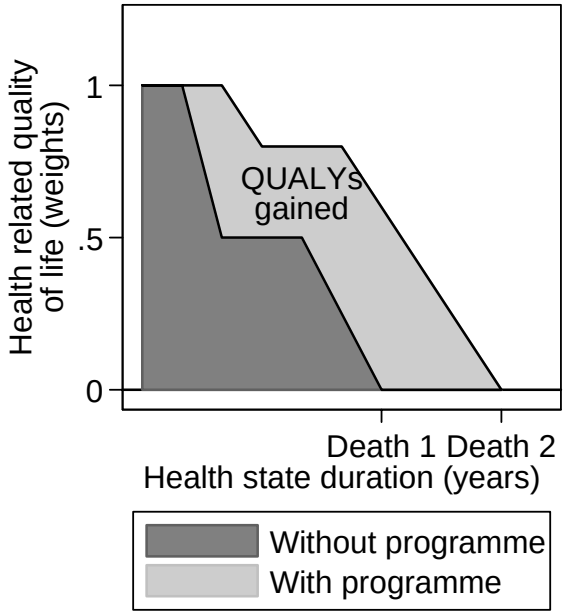
The SF36 has eight domains, i.e. it produces eight different sub-scales:

- Physical functioning
- Social functioning
- Role physical
- Role emotional
- Mental health
- Vitality
- Bodily pain
- General health

The eight domains can be collapsed into two domains: Physical and Mental Health. The advantages of just two domains include being less likely to have a Type I error by carrying out eight separate significance tests and being more resistant to missing items, because each sub-scale has more items in it.

There is a shorter version, the SF12, which has only 12 questions, scored into the two domains of mental and physical health. The advantages of a shorter questionnaire are that there is less data entry and we may get a higher response rate because the questionnaire is less time-consuming to complete.

Figure 3. Expressing impact of an interventions using QALYs



Utility measures

A problem with both disease specific and generic quality of life measures is that the scales do not have ratio properties. This means that a person who scores 60 on the SF12 is better than someone who scores 30 but not twice as good. This makes it difficult to compare across conditions or use for economic analysis. We therefore need a measure which has this property; such a measure is called a utility measure. It enables us to do the sort of thing shown in Figure 2. Here the quality of life is plotted against time. During a year when quality of life is 0.5, the person can be considered to be living not a full year of life but a year of half-quality life, or half a quality life year. They experience a year of life but only half a quality of life adjusted life year, or QALY. Over that year, the area under the quality of life line would be 0.5, and this are is the QALY experienced, 0.5. Over the whole period of follow-up, the area under the curve represents the total QALYs experienced.

Figure 3 shows two quality of life curves for hypothetical subjects, one on an intervention programme and one without, who is the person in Figure 1. The programme has increased the life span of the subject, but has also increased the quality of life while they are living. The difference between the areas under the two quality of life curves is the number of QALYs gained.

Another issue underpinning the valuation of utility measures is the concept of resource scarcity. Economists assign a value to something if one is willing to pay for it. Life has a value because people are willing to trade an increase risk in death to improve their utility (e.g. North Sea divers have an enhanced salary to compensate them for increased risk of death).

Measuring utility

Utility measures are numbers that represent the strength of an individual's preference for a particular health state under uncertainty.

One way of measuring and valuing utility is through a 'standard gamble'. Typically people are presented with a range of scenarios with assigned probabilities. Would you have surgery for your hip arthritis with a probability of dying of 0.01 or not? The values are varied until people accept surgery and this gives a mortality weight to the quality of life. This can be converted into monetary forms using wage differentials, etc., for risky occupations. There are problems with the standard gamble: no-one understands it and it can produce 'illogical' answers, such as people choosing almost certain death for treatment for a minor illness.

An alternative widely used is a Time Trade Off or TTO. In this approach people are given a scenario such as "Imagine you have 10 years left to live in your current health state, how many of these years would you give up to be in perfect health?" The more that is given up the greater the quality of life gained by treatment. TTO, too, has problems: it is difficult to understand and produces many incorrect or illogical answers.

A third approach to determining utility is willingness to pay (WTP). In this approach people are asked about their willingness to pay for a treatment given an outcome scenario. For example, women were asked their WTP for HRT for the treatment of severe menopausal symptoms. Most were WTP significant amounts (substantially exceeding the cost) for treatment. There are problems with WTP. Usually the answer is £25 whatever you ask or people refuse to give an answer or say an infinite amount. On a practical side if you put WTP questions in your questionnaire you get letters sent to MPs with accusations that it is a wicked plot to 'privatise' the much loved NHS.

WTP is also subject to considerable non-response. A study compared the response rates to two methods of eliciting preference WTP and Willingness to Wait. 15% of participants answered WTW but NOT WTP, whilst only 3% would answer WTP and not WTW (79% answered both).

Conjoint analysis is an approach originally used in environmental economics. Patients are given scenarios of a health care intervention and asked to choose (e.g., you get the operation in a month but have to go to a hospital 100 miles away or get it in 8 months at your local hospital). Patients are then asked to choose between scenarios and a utility can be derived between health care scenarios. A problem with conjoint analysis are that difficult questionnaires can lead to misunderstandings. It is currently an approach that is generating a lot of Ph.D.s.

What is the ideal? We want a simple questionnaire that is sensitive and reliable and produces are utility weight for quality of life. We are not there yet – but there are some questionnaires that try. Several utility measurements are available (e.g., EuroQol, HUI) which claim to produce a ratio scale. Their main disadvantage is they are very insensitive to changes in health status.

Figure 4. The EQ5D utility measure

Mobility

I have no problems in walking about ☐
I have some problems in walking about ☐
I am confined to bed ☐

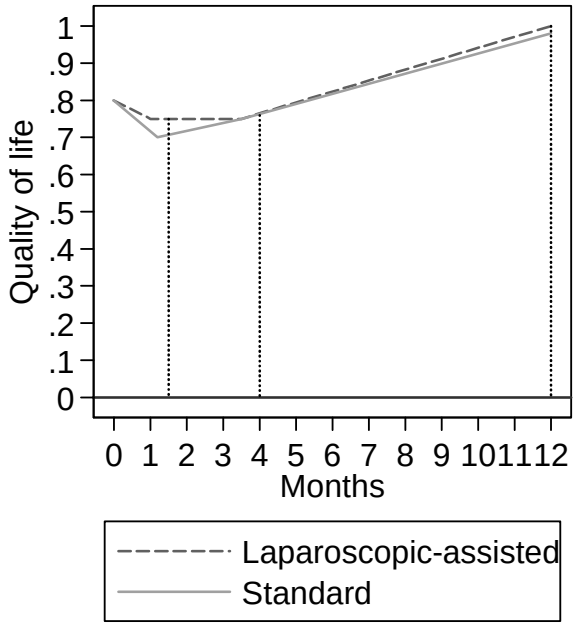
Self-Care

I have no problems with self-care ☐
I have some problems washing or dressing myself ☐
I am unable to wash or dress myself ☐

Usual Activities (e.g. work, study, housework, family or leisure activities)

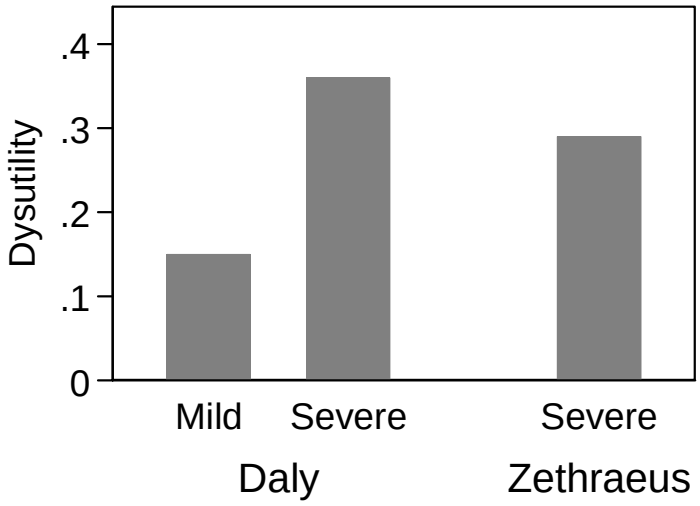
I have no problems with performing my usual activities ☐
I have some problems with performing my usual activities ☐
I am unable to perform my usual activities ☐

Figure 5. Health Outcomes - QALYs



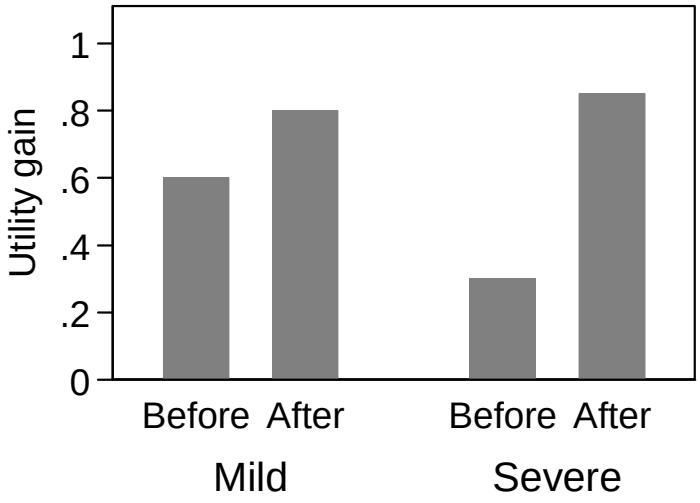
The York local product is the EuroQol or EQ5D. This is shown in Figure 4. The EuroQol scores from 0 (dead) to 1 perfect health. It also allows negative scores (health states worse than death, e.g., a short holiday in Disneyland Paris). There are 245 'health states' in the EuroQol (including negative ones). A large survey using TTO has attached a utility weight to each.

Figure 6. Disutility of menopausal symptoms



(Daly *et al*, *BMJ* 1993;**307**:836-40, Zethraeus, *Health Economics* 1998;**7**:31-8.)

Figure 7. Utility gain of HRT



In figures 6 & 7 the example of the effect of HRT on patients' utility who are on HRT is illustrated. HRT is a very effective treatment for menopausal symptoms and this is reflected in the changes in utility after treatment. However, not all treatments will produce such a large effect on utility measures.

As with all measures and approaches, there are problems with EuroQol. The EuroQol is very insensitive to quite large changes in quality of life. There are large gaps in the scale (e.g., cannot score between 0.88 and 0.99). Figure 5 shows the effect of this, comparing laparoscopic-assisted and standard surgery. Statistically EuroQol has undesirable

properties. It does not follow a Normal distribution, being heavily skewed towards higher values.

Table 1. Some results of the York Back Pain Trial

Scale	Mean change in score		Difference	P value
	Control	Intervention		
Roland & Morris	−1.77	−3.19	1.42	0.02
Aberdeen Back Pain	−8.48	−12.92	4.44	0.01
EQ5D	0.089	0.111	−0.02	0.47

(Klaber-Moffett *et al*, *BMJ* 1999;**319**:279.)

The EuroQol depends on preferences of the sample who completed the TTO survey. Whose preferences should we use? There is debate on how to measure a given health state. Should patients in that health state value it? Or should people not in the health state give it a value. Who to choose matters as generally people in a health state do not value it as badly as people not in a health state. The EuroQol used a general population sample, rather than one selected for having health problems. Roughly there are a couple of arguments for both approaches. Using people who do not have the health condition has the advantage as this is the group of people who are 'paying' to prevent the condition or treat the condition and their 'willingness to pay' values should be used. Conversely those who actually have the condition have greater knowledge of its quality of life effects. The issue is probably less of a problem if evaluations were consistent and all chose the same way of valuing the quality of life loss because generally we are interested in relative comparisons rather than absolute values. Therefore, if we are over valuing quality of life across the board then this does not matter too much if we are merely trying to choose which service or intervention to use A or B. It will matter, however, if it affects budgetary allocations to health care compared with education or other public spending.

Choice of quality of life instruments

What quality of life measures should we use in a clinical trial? Generally, we should use a condition specific measure, a general measure, and a utility measure of quality of life, as well as 'clinical' measures of outcomes.

For example, in the York Back Pain Trial we used the following:

- EuroQol (for economic evaluation);
- Roland & Morris back pain scale;
- Aberdeen back pain scale;
- SF36

Table 1 shows some results. We found significant differences in favour of the intervention in the Roland & Morris and Aberdeen back pain scales but non-significant differences in the EuroQol. Why should this be? One reason may be is that EuroQol is relatively insensitive to changes in small but important measures of outcome. Or alternatively the condition specific measure is too sensitive and is picking up small unimportant quality of life effects.

Economic evaluations want to calculate a cost utility ratio. A EuroQol gain of 0.02, as in Table 1, for a cost of £600 or lower is likely to be cost effective (i.e., <£20,000-£30,000 per QALY, which is a popular view of the NICE cut-off), but difference is nowhere near statistically significant! Is this a Type II error? Possibly as both the 'condition specific measures' showed a significant improvement.

A disadvantage of QALY is its inability to value short intense adverse effects. Consider a local anaesthetic for removing your toenail. Let's assume it costs £5 for the anaesthetic and labour costs. The alternative is for it to be removed without. Assume QoL is 0 for 1 hour after removal. The QALY gain from a local anaesthetic = 0.000114155, divide this into the cost and the ratio = £44,000 – which is not cost effective by the usual criterion! This is nonsense as the WTP for virtually anyone considerably exceeds £5 for an LA.

Generally in trials we use a battery of outcome measures. This means that we can hope to measure all relevant domains. It is also helpful if one is using an 'iffy' outcome measure.

In the VenUS I trial of four-layer versus short stretch bandaging in the treatment of venous leg ulcers, we used the following measures:

- SF12
- EuroQol
- Hyland condition specific measure for leg ulcers.

What did we find? We found that the Hyland measure was very insensitive. It did correlate with ulcer severity but less so than the SF12. There was a lot of missing data from the scale. We, therefore, used the SF12 as the main QoL measure. In other Venus trials (II & III) we are not using the Hyland. In VenUS I, because we had used a relatively sensitive measure of general outcome (SF12) we could still look at Quality of Life.

In the MRC COLPO trial for urinary incontinence the following are used:

- Pad test (jump up and down after drinking a litre of water);
- Bristol Symptom Questionnaire;
- King's Symptom Questionnaire;
- Urinary distress inventory;
- SF36;
- EuroQol;
- Sabbattsberg sexual rating scale.

That's a lot of outcome measures. Why so many? Partly because urinary incontinence affects many aspects of health, but additional urinary incontinence questionnaires were included because referees couldn't agree which one we should use. The principal investigator decided, for a quiet life, to include them all.

Identifying QoL measures

There are huge numbers of QoL measures for nearly every conceivable condition. A website at Oxford has a database of QoL instruments. Often clinicians will develop their own – not generally recommended as it is a specialised task, needing psychologists and statisticians.

What makes a good QoL measure? There are many considerations:

- Appropriateness to the research question.
- Reliability (low measurement error, internal consistency, reproducibility)
- Validity (measures what we want it to measure)
- Responsiveness (changes when quality of life changes)
- Interpretability
- Acceptability
- Feasibility.

There are some statistical properties which we should consider as well. We need a measure to avoid 'ceiling' and 'floor' effects, that is, we do not have respondents starting off at a level where they cannot improve. Some measures have a floor effect and so cannot measure really poor quality of life and vice versa. If a population at baseline either scores nearly the maximum or minimum on a measure the wrong measure is being used.

Choosing an outcome measure

There are a wide number of outcome measures covering most health care states. Before developing a new measure a review (preferably systematic) of the available outcome measures should be undertaken.

For an example, we will consider randomised trial of HRT. It was deemed that a relevant outcome to be measured should be sexual functioning. A systematic review was undertaken to identify a relevant questionnaire.

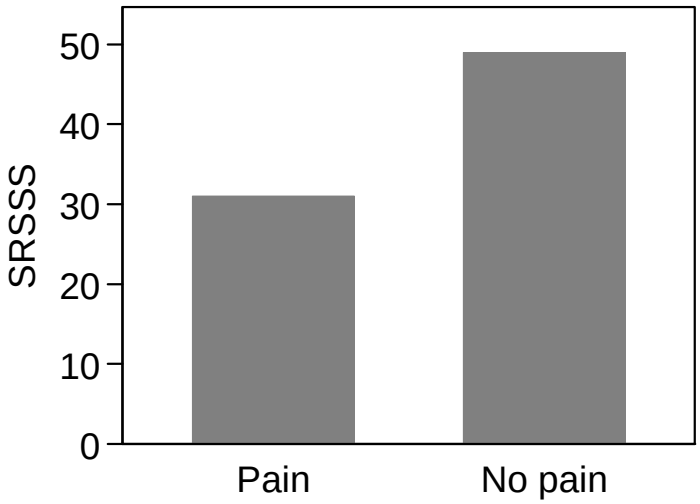
The first issue was relevance to the trial population. Women in the study were 'normal' and were in a HRT study because of low bone density. They were not in the study because of poor sexual health or functioning. Many questionnaires identified in the review were too intrusive for use in this population (e.g., GRISS questionnaire).

One questionnaire the Sabbatsberg Sexual Self-Rating Scale (SSSRS) did not appear to be overly intrusive and was judged to have good face validity by a clinical psychologist working in the field. As far as we could tell, however, it had not been validated properly in any population.

The original questionnaire had 14 questions, but two questions were deemed to be somewhat intrusive and were dropped. The amended version had 12 items. The questionnaire was piloted on an opportunistic sample of women passing through a clinic and these women were happy to answer the questions.

The next stage in the development was to give the questionnaire to 148 women who were being recruited in the RCT. 48 did not respond (35%) of those who answered other QoL questions. As well as the SSSRS we also gave the women the SF36 and the Hospital Anxiety and Depression Scale (HADS). We measure oestrogen levels and sociodemographic variables.

Figure 6. Score differences between those with dyspareunia or not



(Garratt *et al*, *British Journal of Obstetrics & Gynaecology* 1995, **102**: 311.)

We next looked at the number of responses to each possible answer to questions. Each question (item) had 5 possible responses. Guidelines suggest that the endorsement of each item should be less than 80% for it to be valid. This is because if a large majority of people give the same answer, the question cannot distinguish between people and groups. It does not contain much information about the respondent. In our study, the proportion of respondents choosing the most popular item in a question ranged from 36% to 46%. Thus all items were retained for further analysis.

We next asked whether the questionnaire score was correlated with other variables related to sexual response. If the questionnaire was measuring sexual functioning we would expect it to also correlate with other measures of health/wellbeing on the assumption that poor sexual functioning would, on average, have a negative impact on these other domains. As expected the questionnaire correlated in expected directions with seven of the eight SF36 domains and both domains of the HADS. Previous evidence suggests a correlation between sexual functioning and oestrogen levels. The SSSRS correlated with oestrogen levels in the expected direction. We would expect women who experience pain with sex to have lower scores than those who do not. This

was the case with a difference of about one standard deviation. Figure 6 shows the relationship between mean SRSSS score and pain on intercourse.

The score had acceptable statistical properties. Scores tended to have a Normal distribution. This meant that there were no floor or ceiling effects, either of which would make the distribution skew.

We concluded that the SSSRS appeared to be a valid measure of sexual functioning among a group of women aged between 45-49 years.