Many Symptom Scales Do Not Measure Up

Examination of Assessment Methodology

Gunnar Borg, Ph.D., Dr. h.c., Professor Emeritus of Perception and Psychophysics, Stockholm University
gbg@psychology.su.se

Often, people do not seek medical attention because they have a specific disease, but because they do not feel well. This was the underlying meaning of what I was told in the late 1950s by Hans Dalström, the senior physician in Clinical Physiology at Umeå Hospital in Sweden, when he invited me to join a research collaboration on symptoms during exercise tests. We had met previously and discussed how to determine something as subjective as an experience. An experience cannot be measured on a physical scale; rather, it is unspecific and personal and not something objective with a well-defined unit of measurement. Nevertheless, we need to grade our experiences as well as possible: what they are about and how strong they are, both in qualitative and quantitative terms. A quantitative methodology that is generally useful is required, or is it? Is it not enough to just ask a person what he or she is experiencing? Language is, after all, the most important thing for describing something.

The Methodology for Measuring Experiences has a Long History

The methodology for measuring sensory experiences goes far back in time. Simple rating scales with 4-5 multiple choice answers began to be used several hundred years ago. The number of choices was later extended to 7 or more. More were rarely needed for grading a specific event, as was shown in the 1950s [1]. In the 19th century, the physicist and mathematician Fechner formulated an equation for how experiences (R) vary with the intensity of a stimulus (S): $R = k \bullet \log S$ (cf. the decibel concept). However, no new methodology was drawn up. Fechner called this field of research psychophysics. This is not a well-known term. A more understandable term is perceptometry [2]. In the mid 1900s a methodology was developed at Harvard for ratio scaling [3]. It was based on an analogy with a physical scale with a specific unit, equidistant steps and an absolute zero point. Various methods of estimation and production were tested. Magnitude estimation became the most widely used. People were allowed free choice, without restrictions, to assess the strength of their experiences using their own numbers. When almost the entire variation range is used—from very weak to very strong, not just the strength of a single experience—it was found that a very large variation of figures was necessary (0-30 or even larger). Fechner's logarithmic function proved not to be a good match, not even for loudness. Most S-R functions could be described by a simple power function: $R = c \bullet S^n$, in which the exponent according to Stevens [3] ranged from about 0.3 to 3, depending on the sensory modality.

Hans Dahlström and I started using the ratio scaling methodology for perceived exertion during exercise tests. Some patients appeared to overestimate the effort they exerted, such as forestry workers with muscle and joint problems who complained that they had difficulty coping with the heavy piecework. I had studied perceived speed when driving by letting people assess changes in speed during halving and doubling of a vehicle's speed. The results showed that individuals driving at 100 km/hour and slowing down at a 50 sign feel that the speed has decreased by more than half. At just 70 km/h they already think that they are going half as fast. This was then subsequently confirmed by allowing drivers to adjust speeds to what they perceived as half or twice as fast. The perception varied with the square of the speed: $R = c \bullet S^2$. The opinion of some doctors was that forestry workers deliberately exaggerated their efforts to be granted early retirement. Hans Dahlström was of the opinion that this was often not the case. We used my example from the experience of speed to help formulate an explanation. If you experience work as being twice as laborious, a natural consequence could be assumed to be that working capacity is decreased by 50 percent. The S-R function we determined was: $R = a + c \bullet S^n$, where the exponent was 1.6 and an additional parameter, a, had to be added because of the presence of a noise value, a faint sensation of fatigue or aches that were already present during rest. This was one of the first applications of ratio scaling methodology for the determination of symptoms [4, 5].

Problems with Current Symptom Scales

One of the largest and most important areas for symptom assessment is that of pain. Category scales with a fairly small number of multiple choice answers are common, 5 to 7 digits anchored with such expressions as “not at all”, “very weak”, “weak”, “moderate”, “strong,” “very strong” and “maximal”. There are also numeric scales with only anchors (explanatory statements) at the beginning and end. Examples of the first type of category scale are the American scales for angina and dyspnea. The anchors that are selected are not congruent in meaning with the numbers, and differ between the scales, making comparison of symptoms more difficult. Both go from 0 to 4. On the angina scale, 1 is = “light, barely noticeable”, 2 = “moderate” and 3 = “severe'. On the dyspnea scale, 1 is = “mild, noticeable”. “Moderate” can first be found as number 3 [6]. Another example is from the Swedish translation of the International Knee Documentation Committee's (IKDC) form for knee symptoms, in which the various symptoms are determined by different scales [7]. Although they have the same number of scale steps (five), the middle value, “moderate”, is followed by “much” on one scale but by “extremely much” on the other. The same article also provides an example of an 11-point pain rating scale (0 – 10), in which 0 is the “worst imaginable pain” and 10 is “no pain”. This is not good, because an increase in pain is best described with an increasing value. There should also be anchors between 0 and 10. Sometimes there is also a need to have to estimate higher than what the patient—who has a limited experience—believes to be “worst imaginable pain”. The latter is not a good and stable reference level (see below).

The most common pain scale is the Visual Analogue Scale (VAS), which is the standard method in the United States. One of the many problems with VAS is that it occurs in a variety of versions, which causes differences in estimates. The original scale consists of a line with a specific length, such as 10 cm, in which the beginning of the line is marked with “no pain” and the end with “highest (or worst) possible”. The patient can mark with a dash or a slide gage the strength of the pain being experienced. In its original version, VAS works quite well for direct level determinations (despite the uncertainty with regard to the maximum level). Some pain researchers believe that VAS gives a result on a ratio scale [8]. There is, however, no clear support for this opinion, and VAS cannot be found at all in the leading psychophysics textbooks [3, 9]. VAS scales have been designed with the addition of surfaces, e.g. a lying triangle or a line shown as a rectangle cut into pieces and placed after each other so that the continuous variation is destroyed. The surfaces are marked by different colors, numbers from 0 to 10 are added and the line divided with ten dashes. Faces are added, ranging from happy to sad. Words are only placed in rank order, as on a common category scale.

The serious issue with all these additions is that VAS versions vary because the supplements do not increase in congruence with distance and digits. This creates uncertainty, individual differences and technical variation in errors. A patient's pain assessment or results from a clinical study are usually only reported by stating that it was assessed by VAS. But by which VAS scale? In 2012, twelve professors in psychology were asked to assess the magnitude of pain according to the Borg CR100 scale [10]. Six happy to sad faces (as used in several VAS scales) were used. Assessment of pain was considered to be difficult because the faces do not clearly show pain. If they still tried to assess pain, despite the lack of any clear expression of pain (intermodally, as when squeezing a dynamometer), it only began at “extremely weak” at the third to the fourth face. The responses showed that these types of images do not belong on a pain scale. Some scales for quality of life also have shortcomings. One that is commonly used (QOL/Thyroid) has steps from 0 to 10. For physical issues of fatigue and pain, all of them go from “no pain” to “serious pain” (without intermediate anchors). On the other hand, the mental scales have different end anchors. Mental quality of life ranges from “extremely poor” to “excellent” and happiness from “none at all” to “a lot”. It seems sad that feelings of happiness cannot reach higher. An additional example is a scale (EORTC QLQ-C30) that patients use for response after treatment for prostate cancer. Only four steps (not at all, slightly, somewhat, a lot), and not even equidistant steps, are used for questions about whether they have to sit or lie down during the day and if they have trouble taking short walks. A short walk is perhaps 40 meters for one patient and a 4 kilometer round of golf for another.

The RPE Scale for Exertion

The Ratings of Perceived Exertion (RPE) scale from 1970 (slightly modified in 1985) is most frequently used for estimation of a general exertion [11}. It is also called the Borg scale, officially the Borg RPE scale. The CR10 scale is also sometimes used [11]. The RPE scale gives a linear increase with load in an exercise test of steady state. RPE represents a total estimate of exertion, a "Gestalt" (holistic experience) of several integrated symptoms. The two main symptoms in healthy people are breathlessness and muscle fatigue. Slight discomfort from cycling on an uncomfortable saddle, some joint pain, sensations of heat and sensory noise can be added to this. The choice of the Gestalt concept is linked both to my own and colleagues' experiences and to certain definitions of exertion, and from the analysis of a variety of responses from exercise tests. Intra-test reliability is very good (≥0.93 has been noted, parallel test coefficients of the same size and re-test correlations are slightly lower and more variable, 0.75-0.90). High validity correlations between RPE and heart rate (0.70 to 0.85) have been obtained in several surveys. Predictions of working capacity from sub-maximal values are good from the RPE and choice of exercise intensity for fitness exercisers, athletes and patients [11-13]. For patients, the usual breathlessness can turn into shortness of breath, leg fatigue can turn into intermittent claudication, and chest pain can also be added.

Requirements of Rating Scales

There are several requirements that must be addressed in the construction of a good rating scale. The most common is determining the nature of the symptom and its strength "right now". This is often determined verbally. There is a common language in a given country and a given culture and, according to Wittgenstein, a personal language is not possible since language is used for communication. Another requirement concerns determining quantitative strength changes. It is important to know the relationships between different expressions, for example, how much stronger a strong experience is compared to a weak experience. This requires not just an ordinal scale, but also a ratio scale. Adjectives and adverbs can serve as multiplicative constants and be combined with nouns in numerous ways. An interval scale that is equidistant with equal increments is good. However, a scale should preferably also include a zero point, as on a meter scale. This allows you to know that 4 is twice as much as 2 and half of 8. It also allows you to determine a function of the stimulus (S)-response (R) mathematically, for example with the power function: $R = c \bullet S^n$. A high quality scale should also include a schematized conception as an inter-subjective unit. This makes it possible to make extrapolations from sub-maximal determinations to the maximum, for example, aerobic capacity and strength [11, 14]. Occasionally, two extra parameters are required in the power function, a and b: $R = a + c \bullet (S - b)^n$, where a is noise and b an S value that shows the increase of S that is needed for R increase above a. For a healthy person, exertion does not increase (above a possible a value) in a very slow, short walk [11]. Additional requirements concern comparison of symptoms, for example, between leg fatigue and shortness of breath, between pain and exertion, between patients and healthy subjects, and between symptoms and objective findings (for example, between estimated effort and heart rate, and between muscle pain and production of lactic acid).

Stevens placed stringent requirements on the methodology of ratio scaling. It did not work as well as he wanted, but still received the support of several leading psychophysicists, including the expert in theory and mathematics, Luce [15]. This worked well in my first experiments with perceived speed and exertion [11]. A unique study that gave psychophysiological support was between taste and nerve response [16]. One disadvantage of Stevens' method was that it only gave relationships between experiences and not direct levels of experience. That 70 km/hour is perceived as half of 100 km/h is interesting but says nothing about whether 70 is experienced as fast or slow. That depends on the circumstances and the individual. It was regarded as being impossible to develop a valid inter-individual ratio scale that is also level-anchored: "There cannot be a valid unit. This is due to the soul's incurable loneliness," said Professor Ekman at the University of Stockholm. A suggested solution to this "insoluble problem" was my range model (1961 and 1962) with a schematized conception as the unit [5]. An experience of extreme exertion of a maximal character that most people could agree on is the feeling of heaviness when lifting something that is so heavy that it can barely be moved. If this feeling is similar for different individuals of different muscle strength, the subjective variation can be seen as equal, even though it is not objectively equal. A father will perhaps agree with a child who thinks that one stone is twice as heavy as another. But if the child says that the large stone is very heavy, the father may say that it is only quite heavy. The range model provides the opportunity to adjust the measure constant and include two individuals' S-R functions on the same chart (Figure 1). Unfortunately, there is probably no similar simple reference level of inter-subjective validity for pain, taste, nausea or anxiety.

Level Requirements and Ratio Requirements Must be Reconciled

Great demands must be placed on a CR scale (category ratio scale) if it is to reconcile the level requirement with the ratio requirement. The scale must combine linguistic wealth (e.g. "weak" and "strong" in combination with adverbs) with the precision of numbers. The most common CR scale is the CR10 scale®, which is a general symptom scale (Figure 2). It is commonly used for exercise tests and estimation of chest pain, breathlessness and leg fatigue [11, 17]. Although the estimated pain (R_P)

is personal and has great individual variation, some general facts need to be taken into account. A general recommendation for dosage must be based on a certain inter-subjective similarity and then adjusted individually. The ability to adjust the determination can then make use of intermodal comparisons and a schematic unit in the form of inter-individual equivalent exertion (R_{ex}), i.e. $R_p : R_{ex}$. CR10 has also been used for assessment of other symptoms, such as anxiety associated with cardiac rehabilitation, nausea after gastric operations, the need to urinate, and hunger and satiety in anorexia patients. Unlike the RPE scale, CR10 gives a positively increasing function for exertion, i.e. between the linear for heart rate and the positively increasing for lactic acid. The reliability and validity correlations for CR10 are of the same order of size as for RPE [11, 18].

Some of the key requirements are met in this range model: the natural, subjective variation range size, the choice of concise anchors ("precision"), a simplistic schematized main anchor as a unit, the meaning and the positioning of all of the anchors for congruence between numbers and anchors ("interpretation"), a reliable S-R function, avoiding the ceiling effect, and opportunity for the patient to give direct answers and for the therapist to make recommendations on the same scale (the latter does not work with VAS). In addition, there are more common psychometric requirements. Unfortunately, these requirements are not met, and some errors and misunderstandings occur. It should be obvious that the scale and instructions (such as the RPE and the CR10 scales with hundreds of millions of users worldwide) should not be changed. But this is not the case, particularly for the CR10 scale. On occasion the scale is changed for a particular group of patients and for other reasons. For COPD patients, "strong" is changed to "severe" (which complicates comparison of symptoms). Other errors are omitting the half scale step and the opportunity to assess above 10. The instructions are also abbreviated considerably or omitted. Some people can handle this, but many do not. One patient thinks that "be at rest" (before cycling) should be 0, that "moderate" should be about 40 percent of "max" and "strong" 65-70 percent (possibly as in normal distribution). It is important that everyone is familiar with the scale and the instructions before taking the test. A person must also be instructed that they are able to interrupt a test, in accordance with the guidelines in the book "Clinical Exercise Tests" [17].

No rating scale can act as a perfect ratio scale. Even if a scale is the best possible, it must be used by people according to their own opinion. I have often quoted Quine's requirements for sensory evidence and his assertion: "The requirements of intersubjectivity is what make science objective" [1987]. Intersubjectivity can, however, never be perfect. Our experiences differ, as well as our processing and evaluation of impressions, and therefore what we say. Even when we have similar experiences, we can provide different answers. A good scale and clear instructions are prerequisite for reliable assessment of symptoms.

Bad Symptom Scales Should Not Occur

Most doctors are good at identifying symptoms in patient examinations. The method for constructing general scales, however, is poorer. Bad symptom scales should not occur. Several cases of incorrect use have led to incorrect assessment of patients. Different scales for determining the strength of symptoms are not always needed. On the contrary, symptom comparisons are improved when the same scale can be used. The design and use of symptom scales are key quality issues in health care, in both research and clinical work.

Potential ties or competing interests: None declared.

Summary

A review of scales for symptom determination shows that many do not meet the necessary requirements to make them work for direct level determinations and for changes of symptoms and comparison with other symptoms. **Some current** examples are given in which there is a lack of congruence in significance between the figures and so-called anchors in the form of words or images. **This article discusses** several key requirements and how they should be met to provide quality in health care.

Fig. 1
Perceived force of handgrip for two persons squeezing a dynamometer. Solid lines show obtained responses and dashed lines show revision according to the range model. Obtained maximal physical responses (S_1) have been put subjectively equal (R_{tm}).