Measurement Theory for Software Engineers

Although mathematics might be considered the ultimate *abstract* science, it has always been motivated by concerns in the real, physical world. Thus it is not surprising that mathematics includes a branch called *measurement theory*. Some basic definitions from measurement theory can provide a more precise vocabulary for the study of software engineering measurement.

1. Measures and Metrics

Informally, we think of a measure as a way of associating a number, representing some attribute, with a physical object. Such an association is usually called a *mapping* or a *function* in mathematics. More formally, we can give this definition:

Definition 1. Let A be a set of physical or empirical objects. Let B be a set of formal objects, such as numbers. A *measure* μ (the Greek letter "mu") is defined to be a one-to-one mapping $\mu: A \rightarrow B$.

The requirement that the measure be a one-to-one mapping guarantees that every object has a measure, and every object has only one measure. It does *not* require that every number (in set B) be the measure of some object (in set A).

Another term that we use informally in measurement, and one that seems to appear frequently in the literature on software measurement, is *metric*. Some, but not all, measures are metrics. A metric is a way of measuring the distance (itself a term with many interpretations) between two entities, and it has this precise mathematical definition:

Definition 2. Let A be a set of objects, let R be the set of real numbers, and let $m: A \rightarrow R$ be a measure. Then m is a *metric* if and only if it satisfies these three properties:

$$m(x, y) = 0 \text{ for } x = y$$
$$m(x, y) = m(y, x) \text{ for all } x, y$$
$$m(x, z) \leq m(x, y) + m(y, z) \text{ for all } x, y, z$$

We could actually allow other sets of numbers in this definition, as long as the set includes zero and the addition and less-than-or-equal operations are defined on the set.

A common example of a metric is the Euclidean distance metric in the plane. Let $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ be two points in the plane. Define the distance metric *d* by:

$$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Another is the "Manhattan" metric, so named because it is closer to the distance we would travel between two points in (an idealized) New York if constrained to move only along the east-west and north-south streets. For the same two points as above, we define this metric *m* by:

$$m(p_1, p_2) = |x_1 - x_2| + |y_1 - y_2|$$

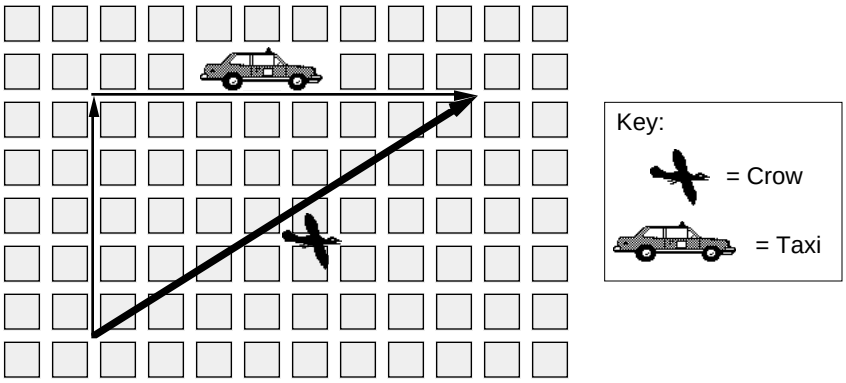

Figure 1. Euclidean distance and Manhattan metrics

Figure 1 illustrates these two metrics. To go five blocks north and eight blocks east "as the crow flies," the distance is approximately 9.43 blocks, but the distance "as the taxi drives" is 13 blocks.

From this definition we see that it is usually imprecise to speak of a "software metric." The preferred term is "software measure."

2. Measures and Relationships

The definition of measure is so general that some measures may not be particularly useful. For example, we can "measure" the members of a football team by the numbers on their jerseys. However, this measure is not very meaningful when used to answer a question like "Does player number 64 deserve twice the salary of player number 32?"

We are all familiar with measures of temperature. Suppose you looked at your Fahrenheit thermometer yesterday at noon and noted that the temperature was 80°, and today at noon it is only 40°. Is today half as warm as yesterday? If the temperature yesterday was only 2° and today it is 8°, is today four times as warm as yesterday?

Suppose your neighbor made the same observations but used a Celsius thermometer. Would the same questions make sense?

Some cities report an air quality index, a measure of the toxic gases and particulate matter in the air. Does an air quality index of 40 mean that the air is twice as unhealthy as an index of 20?

Questions like these assume that two objects have a relationship that can be understood by looking at a relationship between those object's measures. Using the right kind of measure may be critical to answering these questions meaningfully.

For some kinds of objects and measures, the relationships are clear. If one board is two feet long and another is three feet long, the second is longer; furthermore, the relationship between the "amount of board" is the same as the relationship between the measures (the numbers 2 and 3 in this case). Notice also that it is possible to select the longer board without measuring either, which indicates that we use an intuitive notion of "length of board" that is independent of any measure that might be applied.

Sometimes we can manipulate objects to get new objects, and we want to ask questions about the measure of the new object. If one pile contains 50 pounds of sand and another contains 100 pounds, when we combine the piles we expect to have 150 pounds of sand. We implicitly understand that the physical operation of combining piles has the corresponding operation of addition applied to the measures of the piles.

In other cases, the correspondence is anything but clear. Let us consider the *complexity* of a software system. Suppose we are building a software system and we have developed two possible designs. One breaks the system into 10 modules, each with complexity measure in the range of 20 to 30. The other design breaks the system into 20 modules, each with complexity measure in the range 10 to 30. Which design produces a better (less complex) overall system?

This last example illustrates an important aspect of all engineering problems: choosing among alternative solutions. Measurement can be very helpful in such situations, if we use appropriate measures. Most often there are several measures that can be made of the alternatives, and we will need to make *tradeoffs*; we may accept less of one desirable factor to get more of another desirable factor, or we may accept more of an undesirable factor to get more of a desirable one. But how do we use measures to know how much of one factor to trade off for how much of another factor? And how do we use measures to choose among alternatives?

3. Measures and Scales

The problems mentioned above—choosing the longer of two boards or recognizing that combining two piles of sand gives a larger pile of sand—are easy because we have intuitive meanings for the operations of "compare board length" and "combine sand piles." We do not have a corresponding intuitive meaning for the complexity of a software system based on the complexity of its components. We might describe this as an

"intelligence barrier" between questions about objects and the answers to those questions. This barrier is shown in the left side of Figure 2.

Measurement helps us answer these questions as shown in the right side of Figure 2. We first measure the objects in question, yielding (usually) numbers. Then we apply mathematical or statistical techniques to the numbers, yielding another number that somehow relates to the answer to the original question. The final step is *interpretation* of the result, yielding an answer in terms of the original object domain rather than in the domain of numbers.

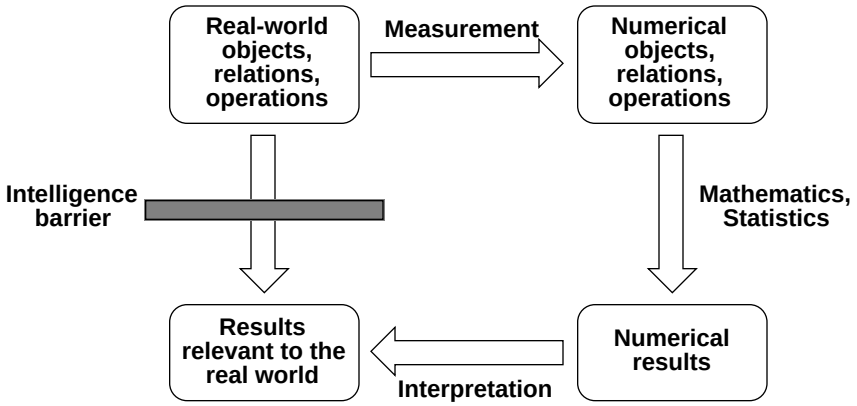

Figure 2. Measurement and the intelligence barrier

For example, suppose you were given a dozen boards of different lengths and asked to select the board of "average" length. It would most likely not be intuitively obvious which board to pick. However, if you measure the length of all the boards, you get a set of numbers. You can then compute the average (*arithmetic mean*) of those numbers. Finally, you interpret this number as the length of the board to be chosen, and pick the board closest to this length.

This technique works because there is an operation on boards that is appropriately modeled by the arithmetic mean operation on numbers. Measurement theory calls this relationship a *scale*.

Mathematically, we can give these definitions:

Definition 3. A *relational system* is defined as an ordered tuple $(S, rel_1, \ldots, rel_n, op_1, \ldots, op_m)$, where:

> S is a nonempty set of objects;
>
> $rel_1, \ldots, rel_n$ are k_i-ary relations on objects in S (this means that the relation rel_i defines a relationship among k_i objects);
>
> $op_1, \ldots, op_m$ are binary operations on objects in S (this means that each operation operates on exactly two objects, producing a third object in S).

An example of a relational system of real-world objects is one with S being the set of all piles of sand, a binary relation "bigger than or same size as," and a binary operation "combine with." An example of a relational system of formal objects (in this case numbers) is one with S being the set of all nonnegative real numbers, the binary relation $\geq$, and the binary operation +.

In fact, these two examples can be shown to be, in some sense, the same if we apply the "weight in pounds" measure to piles of sand, yielding numbers, and we interpret those numbers as weights of piles of sand. Mathematically, we can make this definition:

Definition 4. Let $A = (S_A, relA_1, \ldots, relA_n, opA_1, \ldots, opA_m)$ be a relational system of physical or empirical objects, and let $B = (S_B, relB_1, \ldots, relB_n, opB_1, \ldots, opB_m)$ be a relational system of formal objects (such as numbers). Let $\mu: S_A \rightarrow S_B$ be a measure. Then the triple (A, B, μ) is a *scale* if and only if

$$relA_i(a_{i_1}, \ldots, a_{i_k}) \Leftrightarrow relB_i(\mu(a_{i_1}), \ldots, \mu(a_{i_k}))$$

and

$$\mu(a\ opA_j\ b) = \mu(a)\ opB_j\ \mu(b)$$

for all values of i and j, and for all $a, b, a_{i_1}, \ldots, a_{i_k} \in S_A$.

More informally, this definition says two things. First, for every relation defined on the physical objects, there is a equivalent relation defined on the measures of those objects. By *equivalent*, we mean that if a statement about a relationship between or among objects is true, then the corresponding relationship between or among their measures is also true. Second, for every operation defined on the physical objects, there is a corresponding operation defined on the measures, such that the result of measuring the combined objects is the same as performing the corresponding operation on the measures of the individual objects.

A very mathematical note. The branch of mathematics known as *abstract algebra* deals with abstract entities consisting of sets of objects and associated operations. A mapping from one of these entities to another that preserves the operations in the way described in our definition of scale is called a *homomorphism*. This term comes from the Greek words for *same* (homos) and *form* (morphe). The measure μ in a scale is a homomorphism.

Mathematical measurement theory also helps us classify different kinds of scales and determine whether certain questions can meaningfully be asked and answered about objects measured with different kinds of scales.

To illustrate this, let's return to the discussion question posed earlier about the temperature yesterday and today. You noticed that the temperature yesterday was 80° and today it is 40°; you conclude that yesterday was warmer. Your neighbor, who has a Celsius scale thermometer, noticed it was about 27° yesterday and 4° today; she also concludes that yesterday was warmer.

This example suggests that it is meaningful to make statements such as "Yesterday was warmer than today" regardless of which temperature scale (Fahrenheit or Celsius) we are using. The intuitive concept of "warmer than" is preserved by measurement in the numeric concept "greater than" in both scales.

Now consider the more specific question posed earlier. Suppose you answered "yes" to the question, saying that today (40°) is half as warm as yesterday (80°). But your neighbor with the Celsius scale thermometer observed that today (4.4°) is only about one-sixth as warm as yesterday (26.7°). Who is correct?

This suggests that it is not meaningful to make statements such as "Yesterday was twice as warm as today" because the intuitive concept of "twice as warm" is not accurately reflected in the numeric concept "multiply by 2" applied to temperature measures. Different temperature scales give different results.

What is it about the two scales that make one kind of statement meaningful and another not? The Fahrenheit and Celsius temperature scales are closely related; there are simple algebraic expressions relating a temperature on one scale to a temperature on the other:

$$°F = \frac{9\ °C}{5} + 32 \qquad °C = \frac{5\ (°F - 32)}{9}$$

The relationship between the scales is *linear*, meaning it is of the form $f(x) = ax + b$. This is shown graphically in Figure 3. It is easy to see that what is "warmer" on one scale is also "warmer" on the other scale; this is because the coefficient a in $ax + b$ is positive. But it is also easy to see that "twice as warm" does not have the same meaning on both scales.

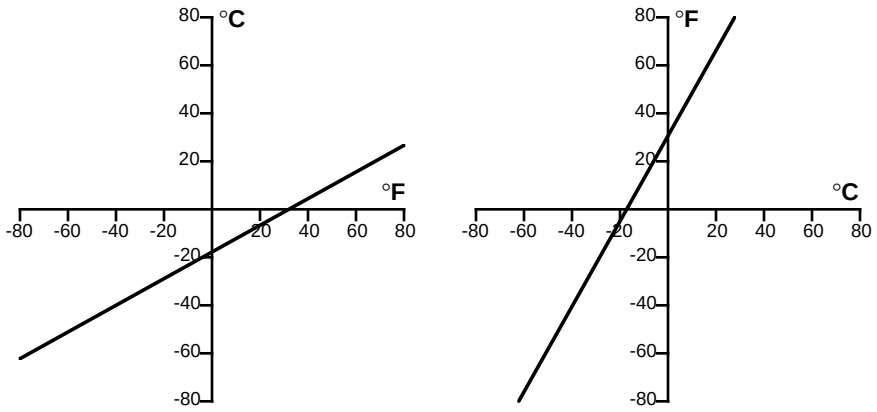

Figure 3. Linear relationships of Fahrenheit and Celsius scales

This example suggests that when two scales are related by a linear function (with a positive coefficient), then "greater than" and "less than" statements should give the same result (true or false) on both scales. A statement can be considered meaningful if it gives the same result on both scales.

Measurement theory allows us to formalize these ideas. First, we can formalize the relationship between two scales of measurement with this definition:

Definition 5. Let (A, B, μ) be a scale, where the set of objects in B is the set of real numbers. Let the notation $\mu(A)$ mean the set of all real numbers that are measures of some object in A. (In mathematics, we call this the *range* of μ.) Then a mapping $t:\mu(A) \rightarrow B$ is defined to be an *admissible transformation* if and only if the triple $(A, B, t \circ \mu)$ is also a scale.

We can interpret this definition as saying that if we have one scale of measure for a certain kind of object, we can invent other, equally good scales by applying admissible transformations to the original scale. Thus if we have the Fahrenheit scale for measuring temperature, we can invent the Celsius scale by applying the transformation $t(x) = 5/9\ x + 160/9$. If we have a scale of length in inches, we can invent a scale in centimeters by applying the transformation $t(x) = 2.54\ x$.

We now have a way of defining the meaningfulness of a statement made about the measures of objects:

Definition 6. Let (A, B, μ) be a scale, where the set of objects in B is the set of real numbers. A statement about the measures $\mu(a)$ of objects in A is *meaningful* if and only if the truth value (whether it is true or false) of that statement is unchanged after applying any admissible transformation to μ.

This definition requires, for example, that any meaningful statement made about the length of an object measured in inches should also be true if the object is measured in centimeters. If we have three boards as shown in Figure 4, then we can make statements such as "Board A is shorter than board B," or "Board B is twice as long as board C." These statements remain true if we measure the boards in centimeters instead of inches.

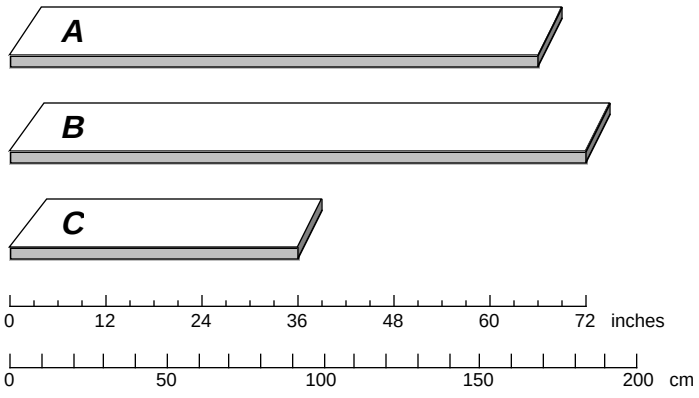

Figure 4. Boards measured in inches and centimeters

4. Classification of Scales

To finish our brief look at measurement theory, we want to consider the classification of scales and the kinds of admissible transformations that exist in each class. Throughout this discussion, we will assume that we are talking about a scale (A, B, μ), where B is the set of real numbers, and transformations t.

Five kinds of scales can be described that are characterized by their admissible transformations:

Nominal scales simply give numeric "names" to objects. (The word *nominal* is derived from the Latin *nomin*, meaning *name*.) Any numbering is as good as any other, so any one-to-one function t is an admissible transformation. We have already mentioned one example of a nominal scale: the jersey numbers of football players. Any numbering of jerseys is as good as any other (except for other considerations, such as the convention that certain numbers represent certain positions on the team, or the fact that a 10-digit number would not fit on all but the very widest of players).

Ordinal scales assign numbers to objects in a particular order, but any numbers that maintain that order are equally good. Any strictly increasing function t is an admissible transformation. An example is the Mohs scale for the hardness of minerals. The original scale assigned, for example, 1 to talc, 7 to quartz, and 10 to diamond. Years later, a revised scale was created that assigned 1 to talc, 8 to quartz, and 15 to diamond. The numbers differed, but the order remained the same.

Interval scales assign numbers to objects in such a way that the interval between two measure values is meaningful throughout the range of values. Only positive linear functions $t(x) = ax + b$ are admissible transformations. We have already seen that the Fahrenheit and Celsius temperature scales are interval scales. A 10-degree difference between 20° and 30° and a 10-degree difference between 70° and 80° both mean the same thing with respect to how much heat is required to raise an object's temperature.

Ratio scales assign values in such a way that the ratio of two measures is meaningful. The only admissible transformations are positive linear functions of the form $t(x) = ax$. Length is a ratio scale, regardless of the unit of measurement, because ratio concepts like "twice as long" are meaningful.

Absolute scales have only one way of measuring objects, and so the only admissible transformation is the identity $t(x) = x$. Counting is the most common example of an absolute scale. Suppose we want to measure the staff size of a software project and make meaningful statements about the staff size. Counting the people is the obvious measure. We cannot imagine a transformation t other than the identity transformation that would make statements like "My project has 5 people on it" and "My project has $t(5)$ people on it" both true for all 5-person projects.

We should notice that this sequence of scales is increasingly restrictive. For example, every ordinal scale is also a nominal scale, but not vice versa. Every interval scale is

also an ordinal scale (and hence a nominal scale), but not vice versa. The relationships among the classes of scales is shown in the Venn diagram in Figure 5.

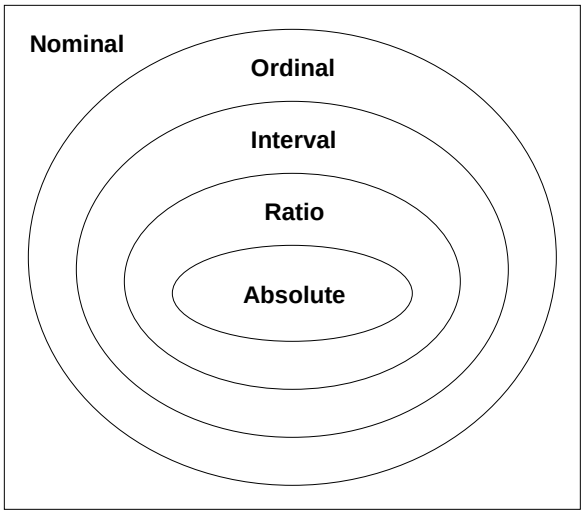

Figure 5. Relationships among classes of scales

5. Applying the Concepts of Measurement Theory

Our brief excursion into measurement theory teaches an important lesson for software engineering measurement: we should consider the kind of measurement scale we must have in order to make meaningful statements about our measurements. For example, if we want to say that one software system is twice as big or ten times as expensive as another, we need to be sure we have ratio scales for size and cost. If we want to talk about the average value of some measurements, we must have at least an interval scale. This lesson can be applied throughout our study of and practice of software engineering.

We can also use these ideas to conclude that it is not meaningful to make the statements "Today is half as warm as yesterday" and "Football player 64 is twice as good as player 32," because neither temperature nor jersey number is a ratio scale.

Discussion Question 1

For each of the following sets of objects, suggest a measure and scale for those objects, and identify the class in which the scale belongs (nominal, ordinal, interval, ratio, absolute).

- Mass of physical objects
- Loudness of sounds
- Brightness of lights
- Human intelligence
- Beauty of the paintings in a museum
- Kelvin scale of temperature
- Size of a software system

Discussion Question 1 (continued)

Productivity of different assembly line workers
Productivity of different software engineers
Cost of different models of automobiles
Reliability of different models of automobiles
Desirability of vacationing in each of the 50 states of the US
Earthquake intensity
Speed of different models of computer
User-friendliness of word-processing or spreadsheet software

Discussion Question 2

The cost of objects is usually regarded as a measure that has a ratio scale; it is meaningful to talk about one automobile model being twice as expensive as another. On the other hand, attributes such as the quality of a car or the complexity of a software system may be measurable only with ordinal scales (or perhaps interval scales). An engineer is often called upon to make judgments in terms of *value*, which we might define as *quality per unit of cost*. For example, should you pay twice as much for twice the quality? Should you pay more or less for software that is more complex? What is "today's best value in a luxury automobile"? When you create a value measure by combining a cost measure on a ratio scale with a quality measure on an ordinal or interval scale, what kind of a scale do you get?

Research Question 3

How does the science of thermodynamics allow us to assert that the Kelvin scale of temperature is a ratio scale and not just an interval scale (like the Fahrenheit and Celsius scales)?