Measurement in the Sciences

If, in a discussion about buying a new table, your spouse were to say to you, “I measured the width of the room and …” you would not expect the conversation to degenerate immediately into a discussion about what is *width*, or what does *measured* mean, or who made your *yardstick*, or what *units* you used. But if, in a discussion with the school guidance counselor, you are told, “I measured the intelligence of your child and …” you could, and probably should, ask those same questions, although they probably won’t be any more warmly received in the guidance office than they were in the dining room.

It’s not simply a distinction between the physical and the social sciences. The distinction is more between the states of the arts than anything fundamental to their essence; the physical sciences have already confronted (surprisingly recently) the basic questions and reached consensus (always more or less provisionally) on what the relevant aspects are and how they are to be measured. In social sciences, we seem stuck in the 19th century and persist in talking about things that aren’t measures and every lord insists on using his own scale.

What Do We Mean When We Say We Have *Measured*?

Suppose at the beginning of the school year, your daughter takes a math test and gets a score of, say, *52*. At the end of the year, she takes a different test and gets a score of, say, *23*. Did she make progress? What if the end-of-year score were *52*? Or *75*?

Here’s almost the same conversation in a different context. On January 1, you *measure* your son’s height and get a measurement of *64*. A few months later, preparing for the start of the new school year, in the course of a physical exam, the nurse measures and reports his height as *173*. Has he grown?

You probably guessed that the first measurement of your son’s height is inches and the second is centimeters. If you must do the math, either multiply the first or divide the second by 2.54 to get both measures in the same units, either centimeters or inches; it doesn’t matter which. If you are doing the back-to-school clothes shopping, you probably don’t need to do the math to know your son has grown four inches or 10 centimeters.

If your daughter is still talking to you, you understand, without the need to understand the test scores, that she knows things at the end of the year that she didn’t know at the start, some of which she learned in math class. While measurement is measurement, there are some important things lacking from the math example that are present in the height example.

First, the <u>units</u> we use to describe height are well established and we have no difficulty converting from one to another and comparing one measure with another. In the school growth problem, we have no idea what the units are. Even I had said the scores are both number correct, or percent correct, or percentile ranks, or stanines, or scale scores, it still would not be clear how the first score compares to the second, regardless of whether you are a psychometrician or a regular parent.

Second, the <u>agents</u> we use to measure height are well established. We do not care much if the first measurement was made with a cheap wooden yardstick given away by the lumberyard and the second with a precisely etched, stainless steel rule intended for use in a physics laboratory. Either instrument is valid enough and reliable enough for the purposes we have in mind here.

The measurement of the height of a person and the distance to Alpha Centauri would be obtained using very different tools but, once we know that the tool was appropriate for the situation, we don't need anything more to interpret and compare the two measures. For the educational measurement, we are paralyzed until we know exactly what the instrument is that was used and even then we are (or should be) rather tentative in our interpretations.

Most importantly, with height, we have a good idea and almost universal agreement about what the aspect is that we are measuring. We talk about it with confidence for the height of people, the width of rooms, the length of homeruns, the size of atoms, or the distance between galaxies. With math, we aren't sure if we are talking about the same aspect for seventh graders that we were for sixth graders.

The math example does not rise to a level that I am willing to call *measurement*.

None of this means that measuring cognitive growth is fundamentally different from or intrinsically more difficult than measuring physical growth. But we are more experienced with the physical, more in agreement about how to go about it, and more comfortable with our results. We understand that physical growth has spurts and plateaus, that the pattern of the events is pretty much the same for every individual, and that individuals vary by months, even years, in the timing of the events. Within some rather loose bounds, we expect and tolerate these individual differences in physical development. Once we understand cognitive development better, it would be very startling indeed if it turned out to be any less interesting or less individualized than physical development.

Fundamental Measurement

L. L. Thurstone, in the 1920's, defined two simple *symmetrical* conditions that must be met to be worthy of his notion of *fundamental measurement*:

- The measurement of the object must be independent of the particular agent used for the measuring.
- The calibration of the agent must be independent of the particular people used for the calibrating.

In our examples, the *objects* are your daughter and son, and the *agents* are the tests and the rulers, the *aspects* are math proficiency and height. Thurstone's first condition means we don't care if you used the wooden yardstick, the stainless steel rule, the Hubble telescope, or a piece of string, so long as the agent was appropriate for the object and the purpose. The second condition means we don't care who else has been measured with the instrument. This effectively eliminates percentiles, stanines, and p-values as possible measures because they compare the individual to some arbitrary group, not to a fixed standard. They are sample-dependent, and thus, paraphrasing Rasch, *scientifically rather uninteresting*.

For Thurstone, *fundamental measurement* was the goal; for Rasch, *specific objectivity* is a reaffirmation of that goal and the method for achieving it. Thirty years after Thurstone told us where to go; Georg Rasch told us how to get there.

The Trouble with Rasch

At the risk of giving away my punch line too early and losing readers who thought perhaps I had been converted, the trouble with Rasch is that it leads to solutions that are too obvious to publish. Hence this undertaking. Most real world measurement problems can be effectively disposed of with counts, sums, and differences. In the nineteen fifties, Georg Rasch (1960) was solving problems manually, often graphically; Douglas (1974) and Wright and Stone (1979) provide some approximate solutions that are almost always close enough and, though perhaps tedious, can be carried out without the benefit of computers. Early calibration programs CALFIT and BICAL (Wright, Mead, and Bell, 1976) ran on a *mini*-computer that would not fit in your dining room and had less memory (by several orders of magnitude) than your cell phone. The simple solutions aren’t what the journals are looking for or the psychometrician guild wants to debate.

Rasch gives solutions that are too simple to publish.

Before we get too complacent about how easy this is going to be, the model requirements are brutal. In its simplest incarnation and slipping perilously close to Metaphysics, the requirement is that *all items, however imperfect, are equally valid and equally reliable* instances of the idea. Success often means painful revisions of our pet theory and abandonment of our favorite agents. Attaining Rasch nirvana does not allow giving up on measurement, applying esoteric mathematics to a more complex model, and settling for *explaining* the data in the barren statistical sense.

Thurstone defined fundamental measurement; Rasch gave us the principles and methods for attaining it, when pursued with sufficient determination. There are three inseparable properties inherent in Rasch measurement, which might be described as the mathematical, the statistical, and the philosophical:

- *Separability*: the parameters are mathematically separate in the statement of the model,
- *Sufficiency*: the parameters have simple sufficient statistics for their estimators, and
- *Specific objectivity*: objective because any appropriate agent will do and specific because no agent is universally appropriate.

Wright (1968) coined the terms “*person-free item calibration*” and “*item-free person measurement*” to describe the two faces of specific objectivity in less philosophically precise language than Rasch used; perhaps Wright feared Rasch’s phrase was beyond his audience. Perhaps he should have said *freed* rather than *free*; he certainly did not mean you can calibrate a test without giving it to some people. Rasch and Wright did mean that it does not matter what people you give it to, within limits. *Within limits* means the people have to interact with the items in a legitimate way. You need agents and a setting that allow the aspect to reveal itself. One would not try to calibrate a yardstick by comparing it to planets or to measure the height of basketball players while they are actively involved in a game; one would not try to calibrate a certifying test for astrophysicists with a sample of fourth graders or undergraduate psychology students.

I would like to add a couple more *S*'s to the standard trilogy: *Simplicity* and *Symmetry*.

The simplicity shows in the arithmetic, algebra, or calculus (depending on your level of mathematical sophistication) but because this is a philosophy book, there is surprisingly little of those activities here. The true test of simplicity is in the application. The essence of the model (the kernel, if you will) is no more complex or less profound (somewhat less monumental in its impact perhaps) than Newton's tour de force: $F = ma$. Estimation of item or equating parameters can be as simple as taking row averages. Control of the model requires recognizing easy items missed by high ability examinees or hard items passed by low ability examinees.

The physical *symmetry* that we all are familiar with has to do with shapes and rotations. The idealized human body has bilateral (or mirror image) symmetry. A snowflake has six axes of symmetry; its shape is invariate to rotations on any of the six. A sphere can be rotated all you want and nothing changes. Mathematicians, of course, have taken this simple idea and generalized it to mean *invariance of important properties to particular transformations*. The question then comes down to what transformations and what properties.

The Rasch model has mathematical symmetry on a couple levels. At the first level, the person parameter and item parameter have identical status in the model's expression. They could be reversed and nothing would change except the signs (and we would be talking about the item anti-difficulty and the person anti-ability.) Equivalently, almost anything we have said, or will say, about the object, we could turn around say about the agent and be equally appropriate.

The invariance that is part and parcel of symmetry also defines Rasch's Specific Objectivity and Thurstone's Fundamental Measurement. The relationship between two objects, between two agents, or between an object and an agent must not be affected by a change in the scaling, to the location on the scale, or the particular company the objects and agents are keeping. A physicist might describe a Rasch analysis as an exploration of the limits of symmetry; Rasch called it controlling the model; I see it as the check for smart people who missed easy items and not smart people who passed hard items.

The other trouble with Rasch, the one that made its distracters decide it doesn't work, is that applying Rasch's methods to build an instrument that conforms to Rasch principles is it's too hard. It requires a theoretical basis for the specific aspect of the object (e.g., person) that we wish to measure and very cleverly and creatively devising agents (e.g., items) that will expose that aspect and a rigorous application of Rasch's methods to control the situation. That's a much higher degree of difficulty than turning the crank to run some data through some canned software.

The Holy Grail of Psychometrics: A Small Illustration of Specific Objectivity

Let's start slow, with a two-item test. After we give it to the 40 students in a school and at a level that we believe appropriate, we might observe the frequency distribution shown in the Hawthorne column of Table I.1. Thinking so far so good, we give our test in another school with 100 students and get the frequencies in the Irving column. You don't need the Rasch model or a psychometrician to see that the Hawthorne students performed better than the Irving students but the table doesn't really tell us how much better nor does it tell us anything about how the items behaved or compare.

Table I.1: Frequencies for a Mythical Two-Item Test in Two Hypothetical Schools

Raw Score	**Hawthorne**	**Irving**
0	8	38
1	16	44
2	16	18

If we get a little more detail by breaking apart the students with raw scores of one, we get the Hawthorne results in Table II.2a, which shows the four rows: Wrong-Wrong, Right-Wrong, Wrong-Right, and Right-Right.

Table II.2a: Detailed Hypothetical Frequencies for Hawthorne

Hawthorne			
Raw Score	Item 1	Item 2	Count
0	Wrong	Wrong	8
1	Right	Wrong	12
1	Wrong	Right	4
2	Right	Right	16

The students who got both items right and the students who got both items wrong tell us nothing about which item is harder; either both items were below the level of the student or both items were above. But comparing the two remaining observations, the number who have *Item 1 right* and *Item 2 wrong* versus the number with *Item 1 wrong* and *Item 2 right*, gives a ratio $D_2 = 12 / 4 = 3$. This is the relative difficulty of item 2 compared to item 1; because it is relative, the difficulty for item 1 can be declared $D_1 = 1$.

Table II.2b: Detailed Hypothetical Frequencies for Irving

Irving			
Raw Score	Item 1	Item 2	*Count*
0	Wrong	Wrong	38
1	Right	Wrong	32
1	Wrong	Right	12
2	Right	Right	18

If we repeat this exercise using the Irving data, we get Table II.2b and the relevant ratio is $D_2 = 32 / 12 = 2.667$, which allowing for rounding errors caused by the small samples and the restriction we only have whole students, is close enough to 3. In spite of differences in the performances in the two schools, the relationship between the two items is the same. This is *person-freed* item calibration.

We will go one step further and take the natural log of the ratios, for reasons not yet disclosed, which gives values of $ln(3) = 1.10$ and $ln(2.667) = 0.98$. We can then say that Item 2 is about one *logit* more difficult than Item 1. (And it does not matter if we declare item 1 to have a logit of zero, ten, or 451; item 2 will always be one logit more difficult.)

This is the simplest illustration of specific objectivity: the two measures of the relative difficulties for the two items were (nearly) identical, although the samples were different students, and the raw score frequency distributions were different, and the proportions correct for each item were different.[1]

Measures for the Two Schools

We are now in a position to compute a measure for each group of students. Because we are in the wonderful world of Rasch, we can base our measures on either item 1 or item 2 and get statistically equivalent results; we will figure out later how to consolidate the information. And using an expression $\frac{B}{D} = \frac{N_{10}+N_{11}}{N_{00}+N_{01}}$, which we will explain later and the data from Table II.2a, the measure for Hawthorne is either:

1. $H_1 = D_1 \frac{12+16}{8+4} = 2.33$, (with a natural log of 0.85,) using item 1 or

2. $H_2 = D_2 \frac{4+16}{8+12} = 2.667$, (with a natural log of 0.98,) using item 2.

In each case, these start with the number of *rights* divided by the number of *wrongs*. The analogous calculations for Irving and Table II.2b are:

3. $I_1 = D_1 \frac{18+32}{12+38} = 1.0$, (natural log = 0.0), or

4. $I_2 = D_2 \frac{18+12}{32+38} = 1.14$, (natural log = 0.13.)

Hawthorne is about *0.85 – 0.0 or 0.98 – 0.13 logits* above Irving, depending on whether item 1 data or item 2 data are used, which isn't bad for estimates based on single items.

I might seem to violating my property of symmetry by not computing the measures for the schools in the same way I did for items. However, we are at a different stage in the process; we used the items to define and anchor the scale (i.e., item 1 is logit 0.) The school measures are calculated on the scale we established with the items. The entire process could be flipped and students used to define and anchor the scale.

Rasch Measurement as Foreign Language

The first Rasch phrase, which underlies all Rasch models, that you will need to use in conversations is:

5. $P_{vi} = \frac{B_v}{B_v + \Delta_i}$, where B_v quantifies the aspect in the object and

Δ_i quantifies the aspect in the agent.

[1] Now for the fine print and disclaimers. These data were simulated with samples one logit apart and items one logit apart. Had we used expected values and the full precision of the spreadsheet program in the calculations rather than whole numbers, both schools would have returned the generating value of 2.718. The corresponding logit would be 1. We have to live with whole students.

This simple expression captures Rasch's requirement that the probability of *success* on a trial is determined solely by the object's power and the agent's resistance. Translating into words that might be used in a school conversation, *the chance that a student will answer an item correctly depends on only two things: how able the student and how hard the item*.

My nomenclature has vacillated between too formal and too casual. I should perhaps provide today's vocabulary list.

Aspect is the property, attribute, characteristic, trait, construct, proficiency, ability, attitude, aptitude, propensity, power, intensity, etc. that we are interested in measuring. It could be the height of a building, mathematical proficiency of a student, mass of a star, mechanical dexterity of a recruit, political leaning of a likely voter, specific gravity of a liquid, hardness of a rock, acidity of a wine, speed of an athlete, and on and on. An aspect is not a thing, but a property of things.

Object is the thing, the thing that contains the aspect we are trying to measure. In social sciences, the object is usually a person in some role, e.g., a student, patient, applicant, candidate, subject, customer, consumer, citizen. Physical sciences are not so restricted; their things can cover a much broader range, e.g., quarks, galaxies, minerals, rooms, tables, whales, or people. An object can have many aspects, e.g., size, height, weight, mass, color, temperature, heat, density, wave length, language fluency, musicianship, political preference, athleticism, sensation of pain, flexibility, but we try very hard to measure one at a time.

Agent is the measuring instrument, e.g., the ruler, test, item, questionnaire, scale, thermometer, odometer, spectrometer, or almost any device whose name ends in *–meter* and many that end in *-scope*. The art of measurement involves devising an agent that will effectively isolate the aspect we are pursuing for the objects that we have.

Ability is the generic term commonly used in the Rasch literature, including this piece of literature, to label the measurement scale for the aspect. It shouldn't be taken to mean any more than that. The term implies nothing about where the person got it, whether it's nature or nurture, which direction is up, if it can be manipulated, or how much is enough. We could be even more erudite and call it β or θ without implying any more or less.

Difficulty is a generic term commonly used to label the scale for the aspect in the item. For educational assessment, it is simply how difficult the item is to respond to correctly, compared to other items. For measuring length, a *difficult* ruler would be one that is longer than the other rulers, making it appropriate for bigger objects. The term makes some sense in educational assessment; not so much in other areas like attitudes or opinions. The distinction between the ability of a person and the difficulty of an item is purely semantic and convention; they are expressed in the same units, may be plotted on the same scale, and subtracted and compared at will[2].

Logit is the native unit of Rasch measurement. It is the *log odds* that the object will best the <u>standard</u> agent; in the language of expression 5, B/Δ are the odds and its natural log is

[2] If we are talking about the same aspect measured with "equated" instruments, which begs a couple questions that we will attempt to dispose of in later chapters. Don't try to subtract and compare degrees Kelvin with degrees Fahrenheit or height with reading proficiency.

the logit. They are not standard normal deviates but they look a lot like them; -4 is a large negative and +4 a large positive. Log odds, or logits, are convenient for doing the arithmetic but probably shouldn't appear in public without some sort of camouflage.

The basic Rasch model is normally written in logits, which emphasizes the importance of the difference between the object and the agent, *β–δ*, which we might call the kernel of the expression:

6. $$P = \frac{\mathrm{B}}{\mathrm{B} + \Delta} = \frac{e^{\beta - \delta}}{e^{\beta - \delta} + 1},$$ where β and δ are the logit versions of the ability and difficulty parameters and equal to the natural logs of B and Δ respectively.

I will generally follow the venerable statistical convention of using the Greek alphabet when I mean the model parameters (e.g., *B, Δ, β, δ*) and the Latin alphabet to mean the estimates of those parameters (e.g., *B, D, b, d*).

Scale Scores are the camouflaged versions of logits, which can make them suitable to be seen in public. They are a linear transformation, which means you start with the logit, then multiple by something and add something. The idea is to make them easier to use, interpret, and remember without losing the interval scale properties of logits. While the somethings you use to multiple and to add can be freely chosen to give whatever scale you like, they are typically chosen to label our two favorite points with nice values.

Scale scores have no inherent meaning. They're just labels; the validity and meaning of the scale do not depend on how the labels are chosen. The point at which water changes from a liquid to a solid is the same point and has the same consequences to us whether we chose to label it 0°C, 32°F, 273° Kelvin, or 492° Rankin. As labels, they have no meaning until we attach meaning. The label 492°R starts to become meaningful when we are told it is the freezing point of water; a temperature of 37°C has meaning in most of the world but doesn't in the US until you know it is the normal temperature of the human body measured orally; a scale score of, say, 1300 may have some meaning if we know this is the mean score of 11th graders in the base year.

Meaning comes from our own experience and through milestones others provide to us. Carefully chosen, the labels we use should facilitate the process of defining aspects and communicating measures. Meaning and communication have more to do with the science, purpose, and audience than with the psychometrician. We could live quite full and happy lives without scale scores.