Exemplifying natural science measurement in the social sciences with Rasch measurement theory

David Andrich
Graduate School of Education, The University of Western Australia, Crawley WA 6009, Australia

E-mail: david.andrich@uwa.edu.au

Abstract. This paper presents a social science example of measurement in which the integer count from a convenient origin is shown to be in equal units. This feature is identical to direct measurement in the natural sciences. The example rests on applying the probabilistic Rasch measurement model to two assessments of the same persons which (i) realises measurement as a real valued comparison between persons invariant with respect to instruments; (ii) characterises random variation of replicated measurements by the Gaussian distribution; (iii) shows that the variance of the replications is the inverse of the size of the unit.

1. Introduction
Although elementary measurement can be understood by school children, it is central to the remarkable advancement of the natural sciences [1]. In principle, measurement is the mapping of an amount of some property of an object onto a line partitioned into equal intervals, called units. The count of units from a convenient origin is the measurement of the object.

2. The abstraction of measurement
The mapping of a property onto a continuous, abstract line joining two points with no magnitude is a deep abstraction. No objects, or their properties, immediately fit such a description. Even a drawn line examined through a microscope shows gaps. Natural objects rarely approximate a line. The elementary measurement of mass with a beam balance requires the control of gravity with three dimensional objects which can be of all shapes and sizes. The measurement of temperature is more complex, both in theory and instrument design. The measurement of the concentration of sugar in juice, where the reading is corrected for temperature, is even more complex, as is the measurement of various electro-magnetic properties. A common feature is reading measurements directly from instruments. Direct reading, termed direct measurement, hides the theory and design which manifest the property measured and controls those properties that might disturb this manifestation.

Despite being ubiquitous, natural scientists seem able to take the common characteristics of measurement for granted, for example, those common in measuring the different properties of mass and temperature of an object – they do not agonise as to what measurement really is. This contrasts with social scientists, both in the successes they have, and in the agonising they do, with measurement.
3. Assuming measurement: Carl Friedrich Gauss (1777 – 1855)

One characteristic of measurement that continues to concern natural scientists is uncertainty, the evidence that either literal or inferred replications of measurements do not give the same measurement [2]. To summarise replications required an intellectual leap to hypothesise that the variation was random, and what were perceived as errors should cancel rather than propagate. Establishing a probability distribution …to describe the distribution of the errors arising in repeated measurement of a fixed quantity by the same procedure under constant conditions…, [3 p1] occupied the best mathematicians of the time and culminated

in the quadratic exponential law of Gauss, \( f_Z(z) = (h / \sqrt{\pi}) \exp(-h^2z^2) \),…which became almost universally regarded in the nineteenth century as “the law of error”. [3 p1]

In the above quote, \( z = x - \mu \) is a real number, \( \mu \) is the mean of the distribution and \( f_Z(z) \) is continuous even though \( x \) is an integer count in the unit of the instrument. In updated notation, \( h = 1 / \sigma \)

where \( \sigma^2 \) is the variance of the distribution. This distribution then provided a criterion that the variation in the measurements was no more than random, with no systematic factor disturbing the instrument. An observation relevant for this paper is that

Gauss commented that (1) (the distribution) cannot represent a law of error in full rigor because it assigns probabilities greater than zero to errors outside the range of possible errors, which in practice always has finite limits; that such a feature is unavoidable because one can never assign limits of error with absolute rigor; but this shortcoming is of no importance in the case of (1), because it “decreases so rapidly, when \( h \) has acquired a considerable magnitude, that it can safely be considered as vanishing.” [4 p2]

This paper takes it that when all of the known or suspected components of error have been evaluated [2], and the variation in replicated measurements is no more than random, then the Gaussian distribution provides a measure of uncertainty rather than of error. The result of the paper accommodates that the range of an instrument is finite, that measurements are discrete integer, unit counts, and that within the range \( x = 0,1,2,...,m \) of measurements the probability of a measurement may vanish. In anticipation of this characterisation, consider the discrete Gaussian

\[
P_{nix} = y_x / \left( \sum_{x=0}^{m} y_{nix} \right) = \{ \exp[-(x - \mu_n)^2 / 2\sigma_n^2] \} / \gamma_{ni} \tag{1}
\]

where \( x = 0,1,2,...,m \) are possible measurements, \( y_{nix} = \{ \exp[-(x - \mu_n)^2 / 2\sigma_n^2] \} / \sqrt{2\pi}\sigma \) is the ordinate of the continuous Gaussian, \( \sigma_n^2 \) is its variance and \( \mu_n \) is the mean of the measurements of object \( n \). The normalisation \( \gamma_{ni} = \sum_{x=0}^{m} y_{nix} \) eliminates \( 1 / \sqrt{2\pi}\sigma \) but retains \( \sigma_n^2 \) in Eq. (1).

4. Problematizing measurement: Georg Rasch (1901 – 1980)

In developing the Gaussian distribution, mathematicians took for granted that they had measurements. In contrast, and arising from his work in the social sciences, the Danish mathematician Rasch problematized measurement itself. In contrast to a representational characterisation which describes measurement as a process in which the operations on the measurements match the possible operations on the properties of the objects, for example amalgamating masses and adding their measurements [5], Rasch abstracted the principle of invariant comparisons: that within a specified frame of reference, the comparison in terms of a real valued number should be invariant with respect to which instruments are used, and vice versa [6]. Andrich [7] shows that this principle has consistent results with other approaches to understanding measurements and that it not only explains deterministic measurement, but in its probabilistic formulation, inherently subsumes distributions of replicated measurements.

If an instrument \( i \) has the measurement range \( x = 0,1,2,...,m \), then the Rasch distribution is

\[
P_{nix} = \{ \exp[(x(m - x) / 2)\Delta_i + x(\beta_n - \delta_i)] \} / \gamma_{ni} \tag{2}
\]
where $P_{ni}$ is the probability of measurement $x$, $\beta_n$ is the measure of object $n$, $\Delta_i = \tau_{i(x+1)} - \tau_{ix} > 0$ where $\tau_{ix}$: $x = 0,1,2,...,m$ are the $m$ thresholds of instrument $i$ which partition the continuum into $m+1$ equal units and where measurement $x$ is the count of the number of thresholds exceeded, $\delta_i = (\sum_{x=1}^{m} \tau_{ix}) / m$ is the mean of the thresholds, and $\gamma_{ni}$ is the normalising constant ensuring equation (1) is a probability distribution [8]. If $\beta_n = \delta_i$, that is $\beta_n$ is at the centre of the instrument, then $E[X_{ni}] = m / 2$, the distributions in equations (1) and (2) are identical and $\Delta_i = 1 / \sigma_i^2$ [9]. Thus the Rasch distribution specialises to the discrete Gaussian and its unit is the inverse of the variance of the complementary continuous Gaussian. Andrich [9] shows an example in which the probability of extreme measurements 0 and $m$ vanish and in which the Rasch and Gaussian distributions remain identical. This paper applies this observation to a real example and shows that in the range in which the Gaussian and Rasch distributions are identical, the integer count can be read directly as a measurement in the unit of the instrument just as it is in the natural sciences.

5. An example of measurement as in the natural sciences

The example involves the assessments of two, three-hour Mathematics examinations of 1093 Year 12 students in Western Australia who competed, together with performances in other disciplines, for entry into universities. In this paper they are referred to simply as Tests A and B. The possible scores of 0 to 100 are integers with a substantial number of scores having zero frequency. To study for Test B, it is necessary to study for Test A, but not the other way around. As a consequence, it is expected that Test B is more difficult and discriminates more, than Test A, that is, its unit is smaller. The two examinations together can be seen as two replications of the assessment of a higher order variable of mathematics knowledge. The syllabuses, the class of persons eligible to take the tests, and the construction and administration of the tests, form part of Rasch’s specified frame of reference.

Table 1 shows estimates ($\hat{\delta}_i$, $\hat{\Delta_i}$) for each test. The conditional maximum likelihood (ML) estimates using the algorithm described in Andrich and Luo [10], are independent of the distribution of the proficiencies $\beta$. It is evident that $\hat{\delta}_A + \hat{\delta}_B = 0$, which is the only identifying constraint required in the estimation. As expected, Test B is more difficult, and has a smaller unit, than Test A.

| Test | $\delta$ | $\Delta$ | $\sigma^2 = 1 / \Delta$ | $\sigma = 1 / \sqrt{\Delta}$ |
|------|----------|----------|--------------------------|-----------------------------|
| A    | -0.245   | 0.033    | 29.989                   | 5.476                       |
| B    | 0.245    | 0.026    | 38.844                   | 6.233                       |

Figure 1 shows the plot of the scores against the estimates of the person proficiencies $\beta$ for each possible integer score on each test. The proficiencies are also ML estimates, with the estimates of the test parameters taken as given. The plot also shows, for each test, the observed means highlighted by dots (●) for 10 class intervals. This graphical test of fit shows the observed and theoretical means are virtually identical. The value $m / 2 = 50$, the centre score of each test, and the range of scores $20 \leq x \leq 80$, are also highlighted. The former is highlighted because it is the point where the Rasch distribution is algebraically identical to the Gaussian, and the latter is highlighted because it is the common range of scores which appear linear with the person estimates. Outside that range, which accounts for the finiteness of the scores, the relationship is curvilinear. However, there is a further justification for highlighting the range $20 \leq x \leq 80$.

Figure 2 shows the Rasch, discrete and continuous Gaussian distributions for two locations for Test A, $\beta = -1.246$, $E[x] = 20$ and $\beta = 0.775$, $E[x] = 80$. Two observations are relevant. First, in the left
distribution, \( E[x] = \mu = 20 \), the probability of the extreme response, \( P_{\text{mix}} = 0, x = 0 \), vanishes, and in the right distribution, \( E[x] = \mu = 80 \), the extreme response \( P_{\text{mix}} = 0, x = 100 \) vanishes. Second, the three distributions overlap, with the two probabilities of the discrete distributions having values of the ordinates of the continuous Gaussian. All distributions in the range \( 20 \leq x \leq 80 \) have these features.

**Figure 1.** Plot of the observed score against the person estimate \( \beta \).

**Figure 2.** Distributions of Test A in which \( P_{\text{mix}} = 0 \) (vanishes), \( x = 0; x = 100 \) respectively.

**Figure 3.** Distributions of Test A in which \( P_{\text{mix}} \neq 0 \), \( x = 0, 100 \) respectively.
In contrast, figure 3 shows the distributions for two locations more extreme for Test A: \( \beta = -1.840, E[x] = \mu = 5 \) and \( \beta = 1.350, E[x] = \mu = 95 \). Neither of the features which hold in figure 2, hold in figure 3: the probability of an extreme response does not vanish and the three distributions diverge. Equivalent distributions for Test B are not shown.

6. Measurement within the range where the Gaussian holds

The paper now focusses on the range of scores \( 20 \leq x \leq 80 \) for both tests within which the discrete Gaussian and Rasch distributions are identical and overlap with the continuous Gaussian, the range in which Gauss’s assumption, that the extreme probabilities vanish, holds. The left panel of figure 4 shows the relationship between the score on each test and the estimated values of the person locations \( \beta \). Both are perfectly linear. This may have been anticipated from figure 1. In addition, the slope of the respective lines is exactly their unit \( \Delta \). Thus for each successive integer score increment, the increment on the estimated variable \( \beta \) is exactly the unit distance \( \Delta \) between the thresholds in the Rasch distribution of equation (2). The right panel of figure 4 shows the linear regression in the prediction of the score on Test A from an integer score on Test B. The slope of this regression line is the ratio of the units of the tests, \( \Delta_B / \Delta_A \). Thus the reading of the integer score on the test in the range \( 20 \leq x \leq 80 \), which is the count of the number thresholds in equal units exceeded, is directly analogous to reading degrees Centigrade on a thermometer, and then the conversion of scores between these two tests is directly analogous to converting degrees Centigrade to degrees Fahrenheit.

![Figure 4](image-url)

Figure 4. Conversion of each test’s measurements to \( \beta \) (left graph); conversion of Test B integer measurement to Test A (right graph).

The Rasch model permits estimating person locations beyond the linear limits shown above, but that topic is not dealt with in this paper, which focusses on an example showing an equivalence between Rasch and natural science measurement.

7. Comment on the abstraction and realities of measurement

These concluding comments consider further the abstraction of measurement, and in particular not being wedded to the literal examples of concatenation considered in representational measurement. This abstraction is consistent with the Bureau of International Weights and Measures redefining the base units, such as the Kilogram, in terms of universal constants rather than an object. However, the requirement of invariant measurements with respect to instruments and the Gaussian as the criterion of uncertainty which is no more than random, remain.

First, \( \Delta_A, \Delta_B \) are estimated average unit distances between successive thresholds, not literally the distance between each pair of thresholds which is the same in their respective instruments. However,
this simply applies the principle of probability distributions to hypothesised randomness, which can be tested using the Rasch model.

Second, the integer scores from 0 to 100 for each test were constructed from multiple items with different maximum scores for each item and the sum of the scores was presented to the nearest integer. Again the principle of randomness in cancelling what may be perceived as errors is implicit.

Third, there is no literal replication of measurements of the same person with the same instrument. In the development of the Gaussian, the number of examples with replicated measurements was small, no larger it seems than 27 [11]. In the above example, there are two measurements from two instruments. From the estimated relative locations and units of each instrument, the Rasch model is the probability distribution of inferred replications for any possible measurement of a person.

Fourth, it is relevant to emphasise the relationship between the measure of uncertainty and the Rasch and Gaussian distributions: the variance is the inverse of the size of the unit, showing that the smaller the unit, which implies greater precision, the greater the variance in the smaller unit.

As indicated in Section 1, measurement of physical properties requires substantial controls and accounting of confounding variables, which in turn is informed by understanding relevant scientific theories. Kuhn [1, p 189] stresses that The road from scientific law to scientific measurement can rarely be travelled in the reverse direction (italics in the original).

In the example of the two measurements of mathematics proficiency, there were also substantial controls which had an eye to something like measurement. The curriculum and syllabuses, which the teachers taught the students who studied for the high stakes examinations, were tightly specified, the exam writers were experienced and wrote questions consistent with the syllabus, the administration of the examination was closely controlled, and the markers were experienced and trained. In this example they also seemed able to produce instruments which measure in a constant unit across a substantial range. This construction and manifestation of a variable is consistent with the kinds of controls on both objects and instruments of measurements in the physical sciences. The difference between those and the above example seems to be primarily in the breadth of the frame of the reference to which the measurement can be applied.

References
[1] Kuhn T S 1961 The function of measurement in modern physical science Isis 52 161-90
[2] JCGM 2008 JCGM 100:2008 Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement (GUM 1995 with minor corrections) (Joint Committee for Guides in Metrology)
[3] Eisenhart C 1983a Law of error I: development of the concept Encyclopedia of Statistical Sciences vol 4 ed S Kotz and NL Johnson (Toronto: Wiley) pp 530–47
[4] Eisenhart C 1983b Law of error II: Development of the concept Encyclopedia of Statistical Sciences vol 4 ed S Kotz and NL Johnson (Toronto: Wiley) pp 547–62
[5] Krantz D H, Luce R D, Suppes P and Tversky A 1971 Foundations of Measurement vol 1 (New York: Academic Press)
[6] Rasch G 1961 On general laws and the meaning of measurement in psychology Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability vol 4 ed J Neyman (Berkeley CA: University of California) pp 321–34
[7] Andrich D 2018 Advances in social measurement: A Rasch measurement theory Perceived Health and Adaptation in Chronic Disease: Stakes and Future Challenge ed F Guillemin et al. (London: Routledge) chapter 7 pp 66–91
[8] Andrich D 1982 A two parameter extension of the Rasch model for ratings Psychometrika 47(1) 105–13
[9] Andrich D 2018 On an identity between the Gaussian and Rasch measurement error distributions: making the role of the instrument explicit J. Phys.: Conf. Ser. (accepted for publication)
[10] Andrich D and Luo G 2003 Conditional estimation in the Rasch model for ordered response categories using principal components J. Appl. Meas. 4 205–21
[11] Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty before 1900* (Cambridge, Massachusetts: The Belknap Press of Harvard University Press)