From Concrete to Abstract in the Measurement of Length

Andreas Stephanou
Senior Research Fellow, Australian Council for Educational Research, 19 Prospect Hill Rd, Camberwell VIC Australia 3124, phone: +61 3 9277 5555
mailto:Stephanou@ACER.edu.au

William P. Fisher, Jr.,
BEAR Center, Graduate School of Education, University of California, Berkeley, CA (USA) & Principal, LivingCapitalMetrics Consulting
mailto:william@livingcapitalmetrics.com

Abstract. The concatenation of units of length is widely viewed as the paradigmatic expression of fundamental measurement. Survey, assessment, and test scores in educational and psychological measurement are often interpreted in ways that assume a concatenation of units to have been established, even though these assumptions are rarely stated or tested. A concatenation model for measurement is shown to be equivalent to a Rasch model: any two units of measurement placed end to end must together be of the same length as either one of them added to itself. This additive principle and a concatenation model of measurement together serve as a heuristic guide for organizing two experimental approaches to calibrating instruments for measuring length. The capacity to reproduce the unit of measurement from theory with no need for repeated empirical calibration experiments, as in the geometrical bisection of the line and the resultant halving of the length measure, is highlighted as essential to demonstrating a thorough understanding of the construct.

1. Introduction
When we measure we compare a property of two objects and express the result in counts of an amount of that property. What do we count? We count repetitions of a unit. Where do we get this unit? It is typically built into an instrument and can be traced to a reference standard. But where did the reference standard come from? It was, of course, instituted at a particular time and place, and was preceded by earlier standard units. History shows us that these earlier units are typically less precise and more variable than later ones. Furthermore, some things became measureable for the first time at specific moments in history [1-3].

How do things become measureable? What was counted before there were units? How do vague comparisons of less and more become focused on differences relative to some unit amount? After all, counts are ambiguous. I may have five rocks, and you may have 15, but I may nonetheless have more rock than you do. How does one overcome the ambiguity of counting numbers? What basis for fair comparisons can be formulated?

These are not just rhetorical questions. Though the natural sciences have long since established reference standard units, psychology and the social sciences continue to grapple with problems of fair comparisons. The natural inclination is, of course, to find something to count, and to use those counts as...
a basis for comparisons of less and more. But just as a higher count of rocks may ultimately amount to
less rock than a lower count, so, too, do we need some way of knowing how to relate counts of correct
answers on tests, or of rating scale responses on surveys, to fair measures of amount. After all, a high
score on a test or survey can result from a person with very low abilities encountering very easy
questions, while a person with much higher abilities could have a much lower score as a result of
encountering very difficult questions.

How can this problem be resolved? Early contributions in this area were made by Thorndike [4],
who raised the question back in 1904:

If one attempts to measure even so simple a thing as spelling, one is hampered by the fact
that there exist no units in which to measure. One may arbitrarily make up a list of words
and observe ability by the number spelled correctly. But if one examines such a list one is
struck by the inequality of the units. All results based on the equality of any one word
with any other are necessarily inaccurate. (p. 7)

Wright [5] shows how the problem of a fair unit has been resolved by a number of different approaches
involving concepts and methods associated with concatenation [6], conjoint additivity [7], statistical
sufficiency [8], infinite divisibility [9], and parameter separation [10]. In every approach, counts provide
a natural place to start, but, being based in concrete events or things in the real world, inevitably must be
replaced by an abstract and ideal unit. Such units are defined in mathematical models specifying the
consistency, homogeneity, monotonicity, etc. to which observations of concrete events must conform
for measurement to be achieved.

The principles informing this distinction between the concrete and the abstract frame the question as
to whether observations of differences in physical extension in space would provide a basis for a
estimating a unit linearly equivalent to centimeters?

More specifically, would the calibrations and measures produced by data conforming to a Rasch
model [5,10] of differences between physically extended objects add up and plot out in a manner
equivalent to the measures of the same objects made with a meterstick? These are the questions
explored in this research.

2. Two experiments in measuring length

Wright [5] describes a basis for additivity in measurement in terms of comparing different sized sticks
each held up against the same sample of persons or objects. The sticks are calibrated to a standard
determined by the consistency of the observed data, and not via reference to a predetermined unit
external to the data. The proportions of marks the sticks receive across the persons or objects indicates
whether they are longer (or shorter), and the persons or objects are measured according to the proportion
of longer to shorter marks they receive across the sticks. The invariant consistency of the comparisons
of the sticks across the persons or objects, and of the persons or objects across the sticks, is of
paramount importance.

This scheme was simplified in two experiments: first, by comparing the lengths of two collections of
different sized sticks, each with the others [11], and second, by placing all the ‘sticks’ on the same edge
of a sheet of paper [12]. In order to imitate the situation of the social sciences, where the precise
difficulty of the items measuring is not known in advance and the differences between them impossible
to make exactly uniform, the units of measurement employed in both of these experiments were
deliberately arranged to vary in size in an uneven fashion. Sometimes the units were clumped together
like questions of similar difficulty, and other times they were spread apart with gaps between them.
In both experiments, the resulting logit measures were correlated and plotted against the corresponding millimetre or centimetre measures of the match sticks or objects. By arranging the study of ordinal observations in this way, the manner in which, and the extent to which, length might be taken as the paradigm of measurement in the social sciences can be tested.

2.1 The first experiment
In the first experiment, a pairwise-comparison method for the measurement of length was implemented by breaking 99 matches in two pieces, each time obtaining a headed match and a simple stick. A random sample of 33 sticks was then selected and used as a set of reference lengths. The 99 headed matches and the 33 reference lengths can be displayed along a scale of length, the headed matches on the left and the reference lengths on the right. When a headed match is at the same location as a reference length, the probability for the comparison to result in 1 (headed match longer than reference length) is 0.5; for reference lengths below the headed match the probability is greater than 0.5 and for reference lengths above it is less than 0.5.

The length of each of the 99 headed matches was compared with the length of each of the 33 reference lengths, recording the result of each of the (99x33) comparisons as 1 (headed match longer than reference length) or 0 (headed match shorter than reference length). In the recorded data, a row of 1s and 0s was formed for each headed match, with each reference stick in its own column.

2.2 The second experiment
In the second experiment, a number of small objects such as books, photographs, picture frames, ashtrays, and a television screen were compared to unevenly spaced lines marked along the edge of a sheet of paper (hash marks) and a score read off. All objects were aligned on one end with the bottom edge of the sheet of paper. The hash marks on the side of the paper were each treated as asking, "Does the object extend from the bottom of the paper to here, or is it longer than this?" Each answer of "longer than" was scored 2, each answer of "to here" was scored 1, and each "not up to" was assigned a 0. A row of 2s, 1s and 0s was then formed for each object, keeping each hash mark in its own column.

2.3 The attenuation paradox
Both experiments resulted in situations exhibiting what is known as the attenuation paradox [13]. Perfect scores, obtained when all observations for a question are in the same extreme category, contribute nothing to an analysis beyond indicating that the test is poorly targeted. Such items are too difficult or too easy for the people tested, or a person is too able or unable for the test. It is as though some objects are shorter than the shortest length unit or longer than the longest, and some units are shorter and others longer than all objects. In either case, there is no evidence available that indicates how much longer, shorter, or too easy, difficult, able or unable the item or person is, and estimation is therefore impossible. There are ways around this problem in education and psychology (for instance, by asking questions that are more closely spaced in difficulty, or that reach up or down into the difficulty range where they are needed), but the measurement of length raises a special problem.

The information available in data of this kind is inadequate for calibrating a new instrument. The problem is that the data exhibit a degree of precision producing a deterministic order more suited to a Guttman scale than to Rasch's model [14]. There is no information about the length between two consecutive reference matches in the data from the first experiment because each reference match provides limited information along the scale. In fact the information it provides is only at its actual length. Any shorter headed match will be found to be shorter, and any longer headed match longer, thus
failing to discriminate between different lengths of headed matches except at its length, within measurement error. Measurement error in this case is very small compared to the values of length that are measured because of the careful nearby comparison of headed matches with reference lengths. If the comparisons of the headed matches with the reference lengths were less precise (e.g., the two sticks were held far away from each other when compared) for each headed match there would be some overlapping of 0s for smaller and 1s for longer reference lengths: the results of the comparisons would be governed by some probability law and the boundary between 1s and 0s would not be as sharp. Figure 1 illustrates the two types of comparisons, the nearby comparisons on the left and the far-away comparisons on the right. These probability curves are for a reference length of 25 mm and give the probability for a headed match to be longer than this reference length. The comparison distance determines the discrimination between shorter and longer reference lengths: the closer the distance, the greater the discrimination. The more gradual slopes to the asymptotes in the Rasch ogive provide information on the relative distances between items, making it possible to estimate their locations relative to one another. This information is absent from the Guttman data, meaning that answers to the questions of how much more each item represents cannot be had.

Because the estimation algorithm will not converge in the absence of overlapping probabilities, a small number of individual observations (less than 1%, 3 cases in the second experiment) were changed to provide the information needed.

2.3 Results
The millimeter values for the 99 match heads units have been plotted against the estimated logits in Figure 2, and centimeter values for the 17 ordinal hash marks in the second experiment are plotted against their corresponding logits in Figure 3. There is a steady progression and one to one correspondence between the meter and logit values that approximates an identity line.

In Figure 3, items 2 and 3 calibrated to the same logit value because there was no object measured that landed between them; there is thus no way of discriminating between them on the basis of these data. Where units 12, 13, 14 and 15 are bunched up within 1.5 centimeters of each other, their logit values are nearly within standard errors of each other. Unit 15 calibrates a little higher than the centimeter values indicate it should have; this is because so few of the objects measured reached it. Overall, the Rasch model places the units of measurement in exactly the same order and with very nearly the same spacing as a meterstick does. Can we say that the units add up the way centimeters do?
Given the linearity of the plots in Figures 2 and 3, with some degree of uncertainty, apparently there is a basis in evidence for the claim that a linear and additive unit of measurement has been defined. What does this mean?

3. Conclusion
Scientific laws assert abstract ideals never observed to hold with perfect precision in real life [15, 16]. Rasch [17] and Box [18] both repeatedly pointed out that scientific and statistical models are not meant to be true, but to be useful.

Structurally invariant additivity exhibited within and across different ways of observing a phenomenon can be scaled, equated, and anchored to a common unit, which is then disseminated for use as the shared language in a field of research and practice. The two different studies reported here reported their results in different log-odds units, but because both conformed to established millimetre and centimetre units, they could easily be equated and anchored to a shared common language.

The results of these two experiments are similar to those seen in other studies of distance, weight, and density [19-21]. In all of these investigations, standard units of physical measurement were reproduced from ordinal observations. If ordinal observations can support the calibration of instruments of physical measurement that demonstrably exhibit conformity to existing unit standards, is there not a strong possibility that ordinal observations might also support the calibration of instruments and the establishment of unit standards for educational, psychological, and social measurement [2]?

References
[1] Heilbron J L 1993 Weighing imponderables and other quantitative science around 1800 (Berkeley, California: University of California Press)
[2] Roche J 1998 The mathematics of measurement: A critical history (London: The Athlone Press)
[3] Zupko R E 1977 British weights and measures: A history from antiquity to the seventeenth century (Madison, Wisconsin, USA: University of Wisconsin Press)
[4] Thorndike E L 1904 An introduction to the theory of mental and social measurements (New York: Teacher's College)
[5] Wright B D 1999 Fundamental measurement for psychology pp. 65-104 in S E Embretson, S L Hershberger (Eds), The new rules of measurement: What every educator and psychologist should know (Hillsdale, New Jersey: Lawrence Erlbaum Associates)
[6] Campbell N R 1920 *Physics, the elements* (Cambridge, England: Cambridge University Press)
[7] Luce R D, Tukey J W 1964 *J Mathematical Psychol* **1** 1-27
[8] Fisher R A 1922 *Philosophical Transactions of the Royal Society of London* **222** 309-368
[9] Kolmogorov A N 1950 *Foundations of the theory of probability* (New York: Chelsea Publishing)
[10] Rasch G 1961 *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **4** 321-333
[11] Stephanou A 2007 *The measurement of conceptual understanding in physics* PhD thesis, School of Physics, University of Melbourne, Victoria, Australia
[12] Fisher W P 1988 *Truth, method, and measurement: The hermeneutic of instrumentation and the Rasch model* Dissertation Abstracts International (Dept. of Education, Division of the Social Sciences, University of Chicago) **49** 0778A
[13] Engelhard G 1993 *Rasch Measurement Transactions* **6** 257 [www.rasch.org/rmt/rmt64.htm](http://www.rasch.org/rmt/rmt64.htm)
[14] Wilson M 2011 *Joint International IMEKO TC1 + TC7 + TC13 Symposium*, August 31st – September 2nd, Jena, Germany, urn:nbn:de:gbv:ilm1-2011imeko-005:8
[15] Butterfield H 1957 *The origins of modern science* (New York: The Free Press)
[16] Cartwright N 1983 *How the laws of physics lie* (Oxford: Clarendon Press)
[17] Rasch G 1973/2011 *Rasch Measurement Transactions* **24** 1309
[18] Box G E P 1979 In R L Launer & G N Wilkinson (Eds.) *Robustness in statistics* pp. 201-235 (New York: Academic Press)
[19] Choi S E 1997 *Rasch Measurement Transactions* **11** 557 [www.rasch.org/rmt/rmt112.htm#Ounces](http://www.rasch.org/rmt/rmt112.htm#Ounces)
[20] Pelton T and Bunderson V 2003 *J App Meas* **4** 269-281
[21] Moulton M 1993 *Rasch Measurement Transactions* **7** 268 [www.rasch.org/rmt/rmt71b.htm](http://www.rasch.org/rmt/rmt71b.htm)
[22] Fisher W P 2009 *Measurement* **42** 1278-1287