The unreasonable effectiveness of theory based instrument calibration in the natural sciences: What can the social sciences learn?

A J Stenner¹, M H Stone² and W P Fisher Jr³

¹ Chief Scientist, MetaMetrics, Inc., Durham, North Carolina, USA
² Professor Emeritus, Department of Psychology, Aurora University, Aurora, Illinois, USA
³ Research Associate, BEAR Center, Graduate School of Education, University of California, Berkeley, California, USA

E-mail: jstenner@lexile.com

Abstract. In his classic paper entitled “The Unreasonable Effectiveness of Mathematics in the Natural Sciences,” Eugene Wigner addresses the question of why the language of Mathematics should prove so remarkably effective in the physical [natural] sciences. He marvels that “the enormous usefulness of mathematics in the natural sciences is something bordering on the mysterious and that there is no rational explanation for it.” We have been similarly struck by the outsized benefits that theory based instrument calibrations convey on the natural sciences, in contrast with the almost universal practice in the social sciences of using data to calibrate instrumentation.

1. Introduction

Why is mathematics so remarkably effective in the natural sciences, and what might the social sciences have to learn from the way it is used in those sciences? Is the effectiveness of mathematics in the natural sciences truly “unreasonable,” as Wigner [1] put it? Previous research on Maxwell’s foundational contributions shows the effectiveness of mathematical model-based reasoning to be rooted in everyday thinking [2], and not in any special capacities associated with scientists or their objects of study. Rasch’s [3] adoption of Maxwell’s method of analogy [2] set the stage for extending the effectiveness of mathematics into the social sciences [4,5].

In our ongoing explorations of the ways the natural sciences and social sciences invoke, define, and engage in measurement, we have identified a number of differences that are not as epistemologically necessary or predetermined as is popularly imagined [6-14]. We have, to some benefit, contrasted human temperature thermometry (Nextemp™ thermometers; see Appendix A) with the testing of

---

⁴ To whom all correspondence should be addressed.
mathematical ability and the measurement of English language reading ability. Although cataloging these differences has been useful, we now believe they are all traceable to a common cause. Physical science measurement virtually without exception takes place in the context of well-developed substantive theory, experimental evidence, and instruments calibrated to uniform unit standards. In the natural sciences, theories are not just compelling stories about the relationships between measurement outcomes (such as the count of cavities turning black on a Nextemp thermometer), unit standards (degrees Celsius), and measurement mechanisms (a chemical specification equation). They are instead sufficiently elaborated and precise in their specifications that they can be used to calibrate instrumentation.

In contrast, throughout the behavioral and social sciences, instrument calibration depends on data, is typically devoid of theory, and is not traceable to a unit standard. We hypothesize that most of the observed differences between behavioral and physical science measurement are traceable to these foundational differences. The absence of theory is the primary determinant of the need for data-based calibration and the lack of efficient methods for defining units and traceability to them. Further, we offer an example of a theory-referenced reading and text measurement system in the educational sciences that exhibits key theoretical, experimental, and instrumentation features analogous to those of human thermometry. Finally, we review the affordances shared by human thermometry and reading measurement [15].

2. A reading example

A consensus unit and systems for ensuring traceability to it are typical of most natural science measurement. Sometimes, as in temperature measurement, the unification process is not fully completed, but for the vast majority of natural science attributes (referred to as constructs in psychometrics), a unification process has resulted in diverse instrument makers sharing a unit of measure even when the measurement mechanisms vary from manufacturer to manufacturer. Mercury in glass tube thermometers for human temperature measurement differ substantively with Nextemp technology, but produce comparable results. Though the measurement mechanisms are drastically different they both report out in either Fahrenheit or Celsius units. In the case of Nextemp thermometry, a chemical specification equation calibrates the instrument in °C or °F. The chemical specification equation derived from experimental evidence enforces the unit, which is embodied in the instrument to a known degree of uncertainty. Similarly, in reading measurement, a text complexity specification equation enforces the unit and ensures that 100L of difference between two readers, two texts or a reader/text encounter is invariant over any of 100+ English reading tests that, at present, employ the unit [14].

Strictly parallel instruments are typical in the natural sciences. Such instruments share a common correspondence table that links a measurement outcome (count of cavities turning black on a Nextemp thermometer) to a °C or °F. The ability to manufacture essentially identical instruments in large quantities is a hallmark of natural science measurement. The specification equation is the recipe for manufacturing and calibrating clones of an instrument. The social sciences borrow the concept and talk about ‘parallel’ instruments or ‘alternate forms’ and advertise that say, form A and B produce exchangeable measures. But without a specification equation it is impossible to manufacture copies or clones that share the same correspondence table. The reading measurement specification equation can be used to build strictly parallel clones of any reading test (see Appendix B). No such capability exists, for example, for mathematics, and this is so precisely because, at present, there exists no specification equation for mathematical ability that can calibrate mathematics test items (see Appendix C). Different
Mathematics tests are empirically linked to a common scale through large scale, expensive field studies typically involving thousands of students.

Typical Rasch model applications in the social sciences are singly prescriptive. The major prescription that data must meet is non-intersecting item characteristic curves (ICCs), which relate the probability of a correct response to the difference between person ability and item difficulty. The data are used to estimate person and item parameters with no a priori constraints on the item parameters. Mathematics ability measurement is achieved in this way, as is typical of much social science measurement. Because there is no strong substantive theory for ‘mathematical ability,’ there is no specification equation and, thus, no potential for theoretically calibrating items/instruments. Instrument calibrations depend on sample data and a property of the Rasch model: when data fit the model differences between persons and differences between items are independent of items and persons, respectively.

Contrast this singly prescriptive measurement framework with the doubly prescriptive models underlying Nextemp human thermometry and the theoretical framework for reading. In both these cases strong substantive theory coupled with either a Guttman model or a causal Rasch model requires not just data fit to the model but also data fit to the theory specified item/instrument calibrations. For Nextemp a chemical specification equation is used as a recipe for the chemical compound that fills each cavity. By precisely varying the amount of additive the difference between any two adjacent cavities in sensitivity to the green component of light is precisely .2 degrees Fahrenheit. The chemical specification equation enforces this common unit difference for each of the 44 adjacent cavity differences across the 90⁰ F operating range for the instrument.

When data fit a doubly prescriptive Rasch model absolute person measures (not merely differences) are independent of items and instruments and are independent of person sample precisely because no person data figures in the instrument calibration process. Theory calibrated Rasch models are, thus, doubly prescriptive: prescriptive as to Rasch model requirements and prescriptive as to the substantive theory i.e. item/instrument calibrations. Person misfit to a doubly prescriptive model signals that the measurement mechanism that transmits variation in the attribute to the measurement outcome (often a count) is not working as intended for that individual. Frequent failures of theoretical invariance forces reexamination of the substantive theory, the measurement mechanism and instrument calibration procedures. Theoretical invariance can be tested within person over time (e.g. reading ability growth trajectories) and when intra individual theoretical invariance holds across persons then inter-individual theoretical invariance necessarily holds i.e. the attribute is homologous [16-20].

Molenaar [18-20] shows that inferences moving in the reverse direction, inferring from inter-individual factor structures something about intra-individual factor structures, is fraught with complications. The fact that so much of social and psychological measurement is based upon factor analysis of inter-individual variation prompted Molenaar [18,20] to call for a Kuhnian revolution, a paradigm shift in the concepts and methods of measurement in psychology. This paper is intended as another in a series of contributions to this revolution [5-15].

3. Conclusion
Unification of measurement refers to a 200-year-old process whereby dozens if not hundreds of distinct scales for measuring a common attribute are, sometimes quickly and more often slowly, reduced to one, two or three exchangeable units of measure. The history of temperature measurement is a paradigmatic case [21,22] that parallels many contemporary measurement movements in the social
and behavioral sciences. Typically, an attribute (construct) captures the imagination of a community of scholars and engineers and different tests, instruments, mechanisms, and scales are proposed for measuring the attribute, and each is uniquely named. Once there is consensus that the selfsame attribute is being measured across these various devices small scale linking studies are undertaken to build conversion tables to re-express one unit in one or more other units. More advanced linking studies reduce the link to an equation °F = °C * 9/5 +32 making for quick and easy conversions. Since at this stage there is often not much to elevate one scale about the competition the market place takes over and ‘unification’, with all its time and cost savings eventually prevails. Sometimes unification is swift and decisive but more often, particularly in the social sciences, metrology is poorly understood and unification plods along.

A useful case study of unification in the social sciences is the longstanding network of reading measures that has linked 100+ English language reading tests across the world, 250,000 book measures and 200 million article measures. The unification process is 27 years old and is accelerating but is far from complete. This effort drew inspiration and strategies from the history of the unification of temperature [13,20,21].

Rather than using factor analysis of inter individual data to define an attribute structure and then asking if this structure obtains when examining intra individual data we suggest the use of substantive theory (in the form of specification / calibration equations) to establish the universality of attribute structure and measurement mechanism at the individual level. Once this is accomplished there is no puzzle about whether between person differences have the same structure as within person differences – of course they do. So, what this analysis reveals is that it is problematic to study between person variation at one point in time to glimpse truths about within person structures over time [8,19]. But the surprise is that if we start with within-person theory-referenced measurement, where in the extreme no two persons have any items in common over 5 years of measurement, then we would not stop for a moment to puzzle about the validity of the claim that at the end of year 1 Jane was higher than Bob but at the end of year 5 Bob was higher than Jane (i.e. a claim about inter-individual variation.) This is yet another benefit of theory based instrument calibration.

Several key features distinguishing physical science and behavioral science measurement systems can be traced to the absence of substantive theory sufficiently developed that said theory can be used to calibrate measurement instruments. Once such a calibration/specification equation is available most of these distinguishing features can quickly and easily imported into the behavioral sciences.

References
[1] Wigner E P 1960 Comm. Pure Appl. Math. 13 1-14
[2] Nersessian N J 2002 Essays in the History and Philosophy of Science and Mathematics ed. D Malament (Lasalle, Illinois: Open Court) 129-66
[3] Rasch G 1960 Probabilistic models for some intelligence and attainment tests (Reprint, Chicago: University of Chicago Press, 1980) (Copenhagen, Denmark: Danmarks Paedagogiske Institut)
[4] Fisher W P Jr and Stenner A J 2013 Pacific Rim Objective Measurement Symposium 2012 Conference Proceedings eds Q Zhang and H Yang (Berlin, Germany: Springer-Verlag) 1-11
[5] Fisher W P Jr 2010 J Phys.: Conf. Series 238 012016
[6] Stenner A J, Smith M 1982 Perceptual and Motor Skills 55 415-26
[7] Stenner A J, Smith M, Burdick D S 1983. J Ed Meas 20 305-16
[8] Williamson G L, Fitzgerald J, Stenner A J 2013 Educational Researcher 42 59-69
The NexTemp Thermometer is a thin, flexible, paddle-shaped plastic strip containing multiple cavities. In the Fahrenheit version, the 45 cavities are arranged in a double matrix at the functioning end of the unit. The columns are spaced 0.2°F intervals covering the range of 96°F to 104.8°F. Each cavity contains a chemical composition comprised of three cholesteric liquid crystal compounds and a varying concentration of a soluble additive. These chemical compositions have discrete and repeatable change of state temperatures consistent with an empirically established formula to produce a series of change-of-state temperatures consistent with the indicated temperature points on the device. The chemicals are fully encapsulated by a clear polymeric film, which allows observation of the physical change but prevents any user contact with the chemicals. When the thermometer is placed in an environment within its measure range, such as 98.6°F (37.0°C), the chemicals in all of the cavities up to and including 98.6°F (37.0°C) change from a liquid crystal to an isotropic clear liquid state. This change of state is accompanied by an optical change that is easily viewed by a user. The green component of white light is reflected from the liquid crystal state but is transmitted through the isotropic liquid state and absorbed by the black background. As a result, those cavities containing compositions with threshold temperatures up to and including 98.6°F (37.0°C) appear black, whereas those with transition temperatures of 98.6°F (37.0°C) and higher continue to appear green" [23]. Thus, the observed outcome is a count of cavities turned black. The measurement mechanism is an encased chemical compound that includes a varying soluble agent that changes optical properties according to
changes in temperature. Amount of soluble agent can be traded off for change in human temperature to hold number of black cavities constant.

**Appendix B. Edsphere**
The Edsphere™ technology for measuring English language reading ability employs computer-generated, four-option, multiple choice cloze items built on the fly for any prose text. Counts correct on these items are converted into reading measures in the standard unit via an applicable Rasch model. Individual cloze items are one-off creations and disposable; an item is used only once. The cloze and foil selection protocol ensures that the correct answer (cloze) and incorrect answers (foils) match the vocabulary demands of the target text. The text complexity measure and the expected spread of the cloze items are given by a proprietary text theory and associated equations. Thus, the observed outcome is a count of correct answers. The measurement mechanism is a text with a specified complexity and an item generation protocol consistent with that text complexity measure. The text complexity measure can be traded off for a change in reading ability to hold constant the number of items answered correctly.

**Appendix C. Mathematics ability measurement**
Mathematics ability measurement consists of a common supplemental metric that locates students relative to a taxonomy of mathematical skills, concepts, and applications. In order to develop the framework, several tasks were undertaken: (1) develop a structure of mathematics that spans the developmental continuum from first grade content through Algebra I, Geometry, and Algebra II content, (2) develop a bank of items that have been field tested, (3) develop the unit scale (multiplier and anchor point) based on the calibrations of the field-test items, (4) validate the measurement of mathematics ability as defined by the framework, and (5) link extant tests of mathematical ability to the scale. The process of scale unification for mathematics ability is well underway [24].

At present the attribute “mathematical ability” is unspecified; i.e. there is no specification equation and associated analyzer that can be used to locate ‘math text’ on the scale. Rather, data intensive methods are employed to calibrate instrumentation and human intensive qualitative analysis is employed to locate math text (e.g. a chapter on adding fractions with uncommon denominators) on the scale. The vast majority of social science attributes are similarly unspecified. By contrasting Nextemp thermometry, theory-based reading measurement, and data-based mathematics ability measurement we hope to illuminate the chasm of difference between instrumentation that employs strong substantive theory and that which does not.

For the vast majority of measurement systems, it is the case “that the difference between any two points for one individual is qualitatively the same as a corresponding difference between two individuals at one time point” [25]; that is, the attribute is homologous. The same cannot be said for many measurement systems used in the social sciences. We propose that the routine adoption of theory based instrument calibrations will pave the way for homologous attributes in the social sciences, thus, assuring that the attribute on which I differ from myself over time is the same attribute on which I differ from my brother [26].