Overcoming the Invisibility of Metrology: A Reading Measurement Network for Education and the Social Sciences

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

A Jackson Stenner
CEO, MetaMetrics, Inc.
Durham, North Carolina, USA
E-mail: jstenner@lexile.com

Abstract. The public and researchers in psychology and the social sciences are largely unaware of the huge resources invested in metrology and standards in science and commerce, for understandable reasons, but with unfortunate consequences. Measurement quality varies widely in fields lacking uniform standards, making it impossible to coordinate local behaviours and decisions in tune with individually observed instrument readings. However, recent developments in reading measurement have effectively instituted metrological traceability methods within elementary and secondary English and Spanish language reading education in the U.S., Canada, Mexico, and Australia. Given established patterns in the history of science, it may be reasonable to expect that widespread routine reproduction of controlled effects expressed in uniform units in the social sciences may lead to significant developments in theory and practice.

1. Introduction

Metrology is the means by which a common frame of reference is provided for measurement applications in widely separated industrial, scientific, and practical tasks. Large volumes of resources—comprising significant fractions of many nations’ economic productivity—are invested in ensuring traceability to standards for various units of measurement. Ironically and paradoxically, the value of the return on these investments is almost entirely dependent on their transparency and the invisibility of the processes by which they were made transparent.

1.1. Transparent measures, invisible processes

Metrology is doing its job best when no one knows it is there. Experimental and theoretical scientists often take little interest in instrumentation beyond the satisfaction of their immediate needs for precision measures. It is not surprising, then, to also find that the intensive work of establishing traceability is largely invisible to the general public and to researchers in psychology and the social sciences, with significant consequences [1-4].

The uniformity of the various phenomena described by natural laws allows scientists themselves, moreover, the convenient efficiency of not needing to specify scale units in statements of laws. This conveniently supports a division of labour in science that separates theoretical work from the calibration of instruments and both of these from the use of theory and instruments in experiments [5].
But the separation of theoretical, experimental and instrument-focused communities has its drawbacks, too. The lack of metrological awareness among theoretical and experimental scientists is especially problematic. Not knowing when or how reference standard units are established reinforces historically long standing assumptions—such as the idea that the universe or nature is inherently and innately numerical, quantitative, or mathematical—that rarely become explicit objects of attention.

The effect of these presuppositions is significant. Huge social, industrial, and economic efficiencies are gained by universal consensus on the facts of complex phenomena like electricity, temperature, and time. Making quantities seem natural is a cultural achievement of the highest order.

But when the undeniably historic and historical quantitative status of scientific objects becomes reified as unquestioned and unquestionable, then the advancement of science must inevitably be put at risk. The two questions few are asking are: (1) how did the natural sciences succeed in making some things seem so thoroughly natural [1-2,6-8], and (2) how might the social sciences learn from those successes? It will be shown that recent advances in reading measurement embody important lessons in this regard for the social sciences.

1.2. Shortsightedly focusing attention on the local measurement outcome

Historians and philosophers of science have been known to systematically cut the technical infrastructure of measurement out of the picture, sometimes literally, as when a woodcut of a laboratory scene printed in its entirety in one place is trimmed in a later publication to exclude the means by which the effect was produced [9]. Transparency in measurement is a two-edged sword. Widely accessible transparent meaningfulness is achieved only to the extent that technical complexities can be ignored. This point was emphasized by Whitehead [10], who observed that “Civilization advances by extending the number of important operations which we can perform without thinking about them” (p. 61). But what happens when those who make those advances do not record—or do not themselves fully understand—how they extended the number of important operations that can be performed by persons unversed in their technicalities?

This general and very difficult problem was noted by Husserl [11]. In his philosophical consideration of the geometric assumptions informing Galileo, Husserl was sensitive to the ways in which an unexamined hidden agenda was allowed to set priorities. Thus we find ourselves in a situation, in accord with the philosophical problems attending the foundations of measurement, in general, where

Metrology has not often been granted much historical significance. ... Intellectualist condescension distracts our attention from these everyday practices, from their technical staff, and from the work which makes results count outside laboratory walls [3].

Researchers in the natural sciences make use of commercially available precision tools calibrated to universally uniform reference standards, standards capitalizing on the value of invariant laws. The transparency of the substantive qualitative meanings shared in an elaborated metrological network renders the effects of metrology's lack of historical significance relatively harmless. There is enough general awareness of the role metrology plays in science and commerce that standards engineering institutions and traditions with the express purposes of enabling standards are in place and can be accessed when they are needed.

2. Metrological traceability for reading measurement

2.1. Consequences for psychology and the social sciences

But in the social sciences, the lack of metrological institutions, methods, and traditions has been catastrophic. As social scientists have long recognized for themselves [12-13], mainstream research methods and statistical models employ scale-dependent ordinal data in a search for a kind of significance that is often irrelevant to and even antithetical to the production of new knowledge. Even when regularities akin to natural laws are sought and found in psychological and social phenomena [14-15], the results are typically assessed in the language and methods of statistics rather than of
measurement and metrology. The human, social, economic, and scientific consequences of this can be profound.

The lack of institutions and traditions concerning metrological traceability and standards in psychology and the social sciences may have more to do with broad and deep cultural presuppositions than with an actual lack of a basis for them in evidence. After all, what systematic program of experimental evaluation has ever irrefutably established that uniform metrics based in lawful regularities are impossible in psychology and the social sciences? Evidence that has been sought indicates that provisional possibilities exist in some circumstances [14-15].

3. Metrological traceability for reading measurement

Educators have long understood the need to provide students with reading challenges appropriate to their reading abilities. The problem of matching individual readers to texts is usually approached in terms of general curricular structures, and teacher training and experience. Theory has not been a focus of primary interest [16-17]. Rasch’s development of a new class of measurement models in the 1950s was an important step forward in improving the quantification of reading ability [18]. An unsought advantage of this research was improved precision in the matching of readers to text.

When Rasch’s concept of specific objectivity obtained in local measures was combined with a general predictive theory of English text complexity in the 1980s, following the work of Stenner and colleagues [19-21], the stage was set for the efficient creation of a network of reading measurement instruments calibrated in a common unit. By the late 1990s, all of the major English reading tests had been brought into the system. These are today complemented by the hundreds of thousands of books, tens of millions of short articles and hundreds of millions of readers that have been brought into the system.

Reader abilities and text complexities are measured in the same unit. The scale ranges from below 200 for beginning readers to over 1600 for very high level readers and texts. Knowing the text measure of a book and the reader’s measure predicts the degree to which the book will be comprehensible to the student. When the two measures match, a 75% comprehension level is expected. In other words, as has been repeatedly verified in experimental tests, a student with a measure of 500 would be expected to answer correctly 75% of the questions on an assessment made from that text.

More than 30 million measures annually are reported in the U.S. from state and commercial assessments, and from classroom reading program assessments, in a common unit of measurement [21]. The 21 state departments of education that have formally adopted this unit for use statewide are shown in the map in Figure 1.

Traceability to the common unit is determined via both empirical, data-based equating studies and theory-based text analyses [21].

Additional features of the system include electronic tools integrating instruction and assessment for mass customized diagnostics [22], and others charting growth in reading ability relative to college and career readiness [23-24]. Establishing this network of comparable assessments required formal relationships with book and test publishers, teachers, schools and school districts, state departments of education, and psychometric researchers. Furthermore, a new array of material practices was needed to give all the parties involved ongoing and verifiable confidence in the theory. Though great efficiencies stood to be gained, credibility demanded a cautious approach to their implementation. Formal documentation of the birth of the traceability system would be a valuable contribution to the sociotechnical qualities of education.
4. Implications for psychology and the social sciences

In 1965, the National Academy of Sciences published a report articulating common assumptions as to the sequence of events supposed to take place in the development of new instrumentation [25]. Four stages were identified:

1. discovery of suitable means of observing some phenomenon,
2. exploration of this phenomenon with special, homemade instruments or commercial prototypes,
3. widespread use of commercial instruments,
4. routine applications of the instrument to control industrial production as well as research."

Textbook assumptions and presentations of this sequence have indoctrinated researchers in the human sciences to believe, mistakenly, that this is the normal sequence of events, and we have indeed largely succeeded in imposing this sequence on ourselves, against our own interests, and against the grain of actual events in the natural sciences. Rabkin [25] points out that

this scheme seems to be at variance with much of the evidence in the history of science. It has been shown that the integration of instruments has been rarely due to the demand on the part of the researcher. Rather it occurs through vigorous supply of advanced instruments on the part of the industry. The company that proposes these four stages in the report has itself had experience when stages 3 and 4 occur in the reverse order and, moreover, stage 4 is by far the most decisive factor in the development of new instrumentation.

The “vigorous supply of advanced instruments”, and not demand, also characterizes the introduction of popular electronic appliances. Just as Rabkin points out has been the case in research, there was little or no clamour among the public for telephones, televisions, faxes, the internet, blenders, or cell phones before they were developed and introduced.

Scientists and the public both tend to think of instrumentation only as tools employed in the service of the individuals who use them. This perspective is at odds with the historical evidence as well as with philosophers’ observations, such as, for instance, Thoreau’s realization that humanity has become the
tool of its tools [26] and Nietzsche’s insight that the victory of science is better cast as a victory of method over science [27].

This alternative perspective is important because, in the history of science, theory follows from extensive experience with instruments more often than instruments are designed and built from theoretical projections. Standardized and commercially available instrumentation make possible the predictable and routine reproduction of scientific effects essential to the conduct of controlled experiments—and so also to the development of precise and accurate theoretical predictions. As stated by Price,

Historically, we have almost no examples of an increase in understanding being applied to make new advances in technical competence, but we have many cases of advances in technology being puzzled out by theoreticians and resulting in the advancement of knowledge. It is not just a clever historical aphorism, but a general truth, that 'thermodynamics owes much more to the steam engine than ever the steam engine owed to thermodynamics. ...historically the arrow of causality is largely from the technology to the science [28].

In the context of reading measurement, the repeated reproduction of consistent results following the work of Rasch and others led to the Anchor Test Study in the 1970s [29]. This study equated seven major reading tests in the U.S. and involved over 350,000 students in all 50 states. But the purely empirical basis of the calibration and the lack of predictive theory meant that the value of the common unit of measurement was lost as soon as new items were added to the tests, which was immediately.

A plain feature of the equated test results, however, was the similarity of the items from different tests that calibrated in the same locations. More common words and shorter sentences characterized the texts associated with easier questions calibrating at the bottom of the scale, in contrast with the rarer words and longer sentences in the texts with the harder questions at the top of the scale. The stability of this phenomenon will come as no surprise to anyone able to read, but its practical application in a predictive theory relating text complexity, comprehension rates, and reading ability was difficult to achieve [30].

5. Conclusion

Historians of science have repeatedly documented the roles in theory development played by researchers with hands-on experience with instrumentation, as when Kuhn [31] notes that seven of the nine pioneers in quantifying energy conversion process were either trained as engineers or were working with engines when they made their contributions. Indeed, this attitude that an instrument can make a science was taken from physics into economics by both Stanley Jevons and Irving Fisher in their uses of the balance scale as a model of market equilibrium [32-33].

Perhaps new instances of this phenomenon will be created by widespread use of Rasch’s models to calibrate multiple instruments measuring the same things. Consistent behaviour on the part of these instruments may lead in due course to clearer understandings of the constructs involved, and so also to developments in predictive theory and more efficient implementations. These positive outcomes may be made more likely by recent work emphasizing the qualitative progression of learning defined by item hierarchies in educational assessments generally [34].

6. References

[1] Latour B 1987 Science in action: How to follow scientists and engineers through society (New York: Cambridge University Press)
[2] Latour B 2005 Reassembling the social: An introduction to Actor-Network-Theory (Oxford, England: Oxford University Press)
[3] Schaffter S 1992 In R Bud & S E Cozzens (Eds.), Invisible connections: Instruments, institutions, and science pp. 23-56 (Bellingham, Washington: SPIE Optical Engineering Press)
[4] Wise M N 1995 *The values of precision* (Princeton, New Jersey: Princeton University Press)
[5] Galison P 1999 In M. Biagioli (Ed.) *The science studies reader* pp. 137-160 (New York: Routledge)
[6] Dear P 2012 *Osiris* 27 37-55
[7] O'Connell J 1993 *Social Studies of Science* 23 129-173
[8] Sundberg M 2011 *Social Studies of Science* 41 107-125
[9] Shapin S 1989 *American Scientist* 77 554-563
[10] Whitehead A N 1911 *An introduction to mathematics* (New York: Henry Holt and Co.)
[11] Husserl E 1970 *The crisis of European sciences and transcendental phenomenology* (Evanston, Illinois: Northwestern University Press)
[12] Berkson J 1938 *American Statistical Association Journal* 33 526-536
[13] Salsburg D S 1985 *The American Statistician* 39 220-223
[14] Burdick D S, Stone M H, Stenner A J 2006 *Rasch Measurement Transactions* 20 1059-1060
[15] Fisher W P 2009 *Measurement* 42 1278-1287
[16] Sadoski M, Paivio A 2007 *Scientific Studies of Reading* 11 337-356
[17] Engelhard G 2001 *J Appl Meas* 2 1-26
[18] Rasch G 1960 *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980 (Copenhagen, Denmark: Danmarks Paedagogiske Institut)
[19] Stenner A J, Smith M 1982 *Perceptual and Motor Skills* 55 415-426
[20] Stenner A J, Smith M, Burdick D S 1983. *J Ed Meas* 20 305-316
[21] Stenner A J, Burdick H, Sanford E E, Burdick D S 2006 *J Appl Meas* 7 307-322
[22] Rivero V 2011, June 10 *Edtech Digest*. Retrieved 23 May 2012, from http://edtechedigest.wordpress.com/2011/06/10/interview-jack-stenner-takes-education-beyond-the-metrics/
[23] Williamson G L 2008 *J Advanced Academics* 19 602-632
[24] Williamson G L, Fitzgerald J, Stenner A J 2013 *Educational Researcher* 42 59-69
[25] Rabkin Y M 1992 In R Bud, S E Cozzens (Eds) *Invisible connections: Instruments, institutions, and science* pp. 57-82 (Bellingham, Washington: SPIE Optical Engineering Press)
[26] Thoreau H D 1854 *Walden, or life in the woods* (Boston: Ticknor and Fields)
[27] Nietzsche F 1967 *The will to power* (New York: Vintage)
[28] Price D J 1986 In *Little science, big science--and beyond* pp. 237-253 (New York: Columbia University Press)
[29] Rentz R R, Bashaw W L 1977 *J Ed Meas* 14 161-179
[30] Stenner A J, Stone, M 2010 *J App Meas* 11 244-252
[31] Kuhn T S 1977 *The essential tension: Selected studies in scientific tradition and change* (Chicago, Illinois: University of Chicago Press)
[32] Boumans M 2001 In M S Morgan & J. Klein (Eds.) *The age of economic measurement* pp. 313-344 (Durham, North Carolina: Duke University Press)
[33] Maas H 2001 In M S Morgan & J Klein (Eds.), *The age of economic measurement* pp. 277-302 (Durham, North Carolina: Duke University Press)
[34] Wilson M R 2009 *Journal of Research in Science Teaching* 46 716-730