

Daredevil Barnstorming to the Tipping Point: New Aspirations for the Human Sciences

William P. Fisher, Jr.
MetaMetrics, Inc.

Aviation history provides an apt metaphor for the state of Rasch measurement practice, and its potential future. Flying was initially widely believed to be nothing but a spectacular and dangerous fad. Few saw in it any potential for the huge industry that it is today. The current state of Rasch measurement practice is quite akin to daredevil barnstorming in that the field is focused on isolated demonstrations of disconnected technical effects. Only when the analogues of air traffic control, airports, support staff, training programs, textbooks, and partner industries (hotels, restaurants) are in place will Rasch measurement come into its own as the technical medium of a widespread industry. The point at which current practice tips into a new paradigm depends on the realization of operationally validated theory in a supportive social context. The paper closes with speculations on what crossing Rasch measurement's tipping point might entail.

In the early days of aviation, pilots flew their planes around the countryside, attracting attention with stunts, landing in fields, and charging locals a fee for a ride up into the wild blue yonder. As the technology and familiarity with it advanced, flying became more reliable and public trust grew. But in the beginning, flying was widely considered nothing but a dangerous fad. Few believed it would ever evolve into a routine and global industry serving millions of customers.

Current Rasch measurement practice defines something of an analogous daredevil barnstorming stage in the evolution of measurement technology. Ace Winsteps, RUMM, and ConQuest pilots skim the treetops, fly under bridges, loop-the-loop, and perform other amazing feats of invariance, unidimensionality, and model fit, supported by their necessary and sufficient statistical mechanics. Pilots, crews, and passengers all get off the ground and experience great thrills, but no one actually goes anywhere.

Why not? Well, the most obvious reason is that there aren't any airports, or any of the rest of the air traffic control infrastructure of facilities, standards, and communications needed for coordinating activities across a far-flung network. The fact is, there are (almost) no networks connecting the sites from which Rasch flights take off and land. Successful individual flights give a few thrills, or solve an immediately pressing problem, like crop dusters killing boll weevils in the rural American south's cotton fields in 1920.

So just as was true of the first airplane manufacturers, each Rasch workshop takes advantage of the same scientific principles in developing its technology, but is oblivious to the opportunities that exist for systematically deploying that technology to do its job, connecting people separated by time and space into a community of individuals held together by a common language and practices, and differentiated by their variable skills and imaginations.

The airline industry evolved as public trust and familiarity with avionics grew. That growth was directly related to experience, and the technology's visible utility in World War I. Simi-

larly, as isolated pockets of researchers and scale users gain experience with their measurement constructs, they will eventually find the ways through which their different brands of airplanes will be able to land and take off from the same airports, using universally shared communications protocols and aviation procedures.

There are two likely ways in which we will realize this goal, just as there are two different forms of standards found in the International System of Weights and Measures (referred to as the metric system, or the *Système International*—SI—for short). The only remaining instance of a standard based in a concrete exemplar is that of mass. At the end of the 19th century, the kilogram was defined as the mass of a cubic decimeter of water at sea level, and is now defined by the mass of a platinum-iridium prototype. Every metric weight scale is calibrated via traceability to this reference standard. Though there are as yet no instances of an analogous kind of traceability to one instrument's particular item content accepted as the reference standard definition of a construct in the human sciences, it may be possible to create this kind of a network of interconnected instruments.

The kind of reference standard traceability most commonly found in the metric system is the absolute standard, in which the construct is defined mathematically in terms of lawful proportionate relationships known to hold between variables. This kind of standard makes traceability to the reference standard portable, since anyone with the right equipment can reproduce it at will. In the Rasch context, the realization of absolute standards will take more than the equating of instruments. It will take construct theories capable of accurately predicting new items' calibrations on sight, at the first encounter, in the same way that any well-designed and -piloted plane can take off and land at any airport.

There seems to be a widespread misperception among Rasch measurement practitioners that rigorous measurement theory replaces the need for rigorous construct theory. Many articles and conference presentations fail to say anything about the construct, the rationale structuring the

item hierarchy, or even the item content, seemingly on the basis of the assumption that fit to the model and a reliability coefficient complete the task of instrument calibration. But, on the contrary, the maxim, “there is nothing so practical as a good theory” (Lewin, 1951, p. 169), comes into its fullness of meaning only when the properties of a construct are understood well enough to make the quantitative value of any instance of that construct recognizable for what it is at the point of use by any user.

Airline pilots, for instance, know how to fly planes, but are unlikely to know how to build them, and are even less likely to be able to make generalizations from aerodynamic theory in the manner of an engineer or aircraft designer. Even so, pilots licensed to fly a particular class of planes can handle any plane of that type from any airport suited for it, carrying any group of passengers who number within a certain limit. Students in a pilots’ school learn to fly commercial airliners in particular trainers and simulators, but are licensed to fly any aircraft of a given class. Passengers, in turn, are able to fly from place to place with little or no concern for the particular airline or airport involved, or who the pilot, stewardess, mechanic, or ticket agent is. Each different group of people plays a different role and recounts a different story of its interactions with avionic technology, and so comprises an alternative facet in the “fractionally coherent” narratives of avionic technoscience (Law, 2002). Even when most or all of the relevant audiences find their interests embodied in the technology and they lend their support to its development and deployment, it still may not get off the ground, as Law (2002) shows to have been the case with the TSR2, a British military aircraft.

Similarly, test and survey design not only should conform with established scientific principles, but end users should be provided a standardized technology that unifies the principles and applications of that technology for the entire field. Such innovations might be introduced by one or more of the barnstorming and crop dusting pilots themselves, just as Delta Airlines was born in the 1920s. But what do we have to do to

effect a transformation analogous to that characterizing the emergence of the airline industry? At what point will Rasch measurement cease to be a phenomenon of disconnected barnstorming events and become an essential component of unified industries?

Finding the Tipping Point

To borrow Gladwell’s (2000) terms, who are the Rasch innovators, mavens, connectors, and salespeople who will develop the new ideas, provide the information resources, bring people together, and take it all to market? What is the relevant frame of reference for focusing others’ attention on the most salient aspects of the new ideas? How many times does the basic message have to be repeated before it sticks? Where is the point at which Rasch measurement will be pushed over its tipping point, out of its dazzling but local displays of technical wizardry to more mundane but practical, effective, and interconnected routine applications?

Rasch’s tipping point is in fact well defined as the difference between three cells and the fourth in a two-by-two table showing the overlap between ungeneralized and generalized know-how and know-why (Lapr e and Van Wassenhove, 2002). Ungeneralized know-how combined with ungeneralized know-why is the worst case, and amounts to little more than firefighting, rushing from one crisis to another. Most studies of test, survey, and performance assessment data using true score theory and multi-parameter IRT are here in this cell of the table, since their descriptive orientation ties them to individual groups of items and examinees/respondents. In the firefighting orientation, there is neither any generalized know-how concerning instrument craftsmanship, nor any generalized theory of the variable.

Ungeneralized know-how combined with generalized know-why gives nothing but unvalidated theories; there are lots of ideas and talk, but problems are solved one-by-one, locally. Generalized know-how combined with ungeneralized know-why provides artisan skills; people know how to do things, but they cannot explain why what they do works, and they can-

not apply their techniques outside of a given domain, as they have no effective theory.

Rasch measurement theory and practice generally vacillates between these two cells in the table. Virtually all applications of Rasch's models fall in the artisan skills cell, where research has arrived at a generalized understanding of a variable, but no one understands the variable well enough to write a construct specification equation for it. Here, a plane that flies has been built, but the aerodynamics of lift are not well enough understood for anyone else to build one without copying exactly what the Wright brothers did.

Conversely, virtually all Rasch theory falls into the unvalidated theory cell, where logic has arrived at a simple and beautiful mathematical expression, but no one has shown it to have any generalized practical value. This would be a situation wherein aerodynamics is well understood, the plane flies like a dream on paper, but the plane cannot be built, or it doesn't have any landing gear, or an airport to land at.

The tipping point, the place we are all trying to go, is in the fourth cell, where generalized know-how combines with generalized know-why in operationally validated theories. This is the point at which the variable is understood well enough to automate item writing, where the mathematical regularities comprising variation in the construct are so well understood that theoretical and empirical item calibrations correlate .85 and higher. This is the point at which end users completely unaware of the principles structuring the variable are nonetheless able to employ technically advanced instruments that provide quantitative measures at the point of use in a reference standard metric shared by everyone working in the field, from clinicians to patients to accreditors to payers to administrators to researchers to textbook writers, from teachers to students to parents to principals to researchers to curriculum experts.

How will Rasch measurement surpass the tipping point into the world of operationally validated theories? Plainly, by focusing efforts from within the adjacent cells of artisan skills and unvalidated theories. As has been shown time and

again, well-designed instruments with items written by people experienced with the construct can result not only in highly reliable measures, but close study of the overall meaning shared by all of the items in an empirically calibrated hierarchy can also suggest a theory of the construct (Wright, 1994).

Repeated qualitative observation of a pattern in a process or repeated series of events is in fact a common way in which testable hypotheses concerning a theoretical construct emerge. Sometimes the qualitative patterns are in fact early efforts at quantification, but they might also emerge from focus sessions, conversations, or conceptual analysis. When the construct-relevant know-why is combined with rigorous measurement know-how, research can lead to finely tuned instruments, rich theoretical descriptions, precise mathematical formulations of the absolute standard to which all measures can be referenced, and a new degree of practicality at the point of use.

This is approximately what has been happening in the domains of reading theory (Wright, Stenner, and Vanezky, 1995; Burdick and Stenner, 1996), developmental theory (Dawson, 2002a, 2002b), cognitive theory (Wright and Stone, 1979; Carpenter, et al. 1990; Green and Kleuver, 1992; Green, Kleuver, and Wright, 1994; Embretson, 1996; Stenner and Stone, 2003), and genetics (Markward, 2004; Markward and Fisher, 2004). In each of these areas, rich theoretical understandings of qualitative patterns were combined with strict tests of the invariant generalizability of those patterns, and resulted in what could yet prove to be absolute standards to which all measures might someday be traced.

But elaborating a valid quantitative theory, integrating know-how and know-why in an effective technology, is itself not enough to cross the tipping point, as the example of the British TSR2 military aircraft (Law, 2002) shows. One of the crucial factors in reaching the tipping point is what Gladwell (2000) calls stickiness: the number of times a message has to be repeated before it is remembered. Marketing experts commonly hold that even the simplest advertising message has to be encountered seven times before it is remembered.

An effective way of framing the problem of stickiness is to think of it as requiring an agent like a contagious virus, something that will cause the spread of a social epidemic. What is the infectious agent sticky enough to spread the core idea, the meme or self-reproducing unit of cultural evolution (Dawkins, 1976), of better measurement? It cannot be stressed enough that the answer to this question follows entirely from the way stickiness depends on recognizability. New information is most readily absorbed and retained when expressed in familiar terms, images, and symbols, when it takes up people's own concerns in their language.

Rasch measurement has the special advantage of both embodying and facilitating the process of learning through what we already know, which enables it to tap deeply and widely held presuppositions about the creation of meaning (Fisher, 2003b, 2003c). Though this might seem to ensure that the Rasch meme will eventually come to infect virtually every medium of human relating, the point at which it will cross its tipping point remains an open question. Someone, somewhere, at some time, some group of mavens, connectors, and sales people, is going to translate the Rasch meme by dropping extraneous details and exaggerating others so that its deeper meaning becomes intuitively obvious, accessible, useful, and contagious on a mass scale.

Wide scale infection by the Rasch meme will be built on models, estimation methods, software, fit statistics, education and training programs, textbooks, conferences, research publications, mentoring, applications, instrument equatings, etc., but these alone are insufficient to the task. These are the tools needed by the mavens and their wide-ranging knowledge of everything Rasch; the connectors and their ability to bring people with problems together with those with solutions; and the sales people and their ability to get ideas across.

But also of vital importance is the fact that these people—us—use their tools in the context of a shared community, a haven of safe but critical airing and nurturing of new practices, an incubator for our contagious meme. We often un-

derestimate the importance of this context, but our distributed collective memory produces quite tangible group-level harmonic effects that no one of us can produce alone (Wegner, 1991; Hutchins, 1995). These effects are profoundly and decisively central to scientific advance (Latour, 1995). The extent to which we cooperatively work together to create common methods, tools, and languages defines the extent to which we magnify small effects into large ones. The extent to which we compete for the spotlight, or for special credit with respect to this or that innovation, defines the extent to which we reduce small effects to even smaller ones.

Conclusion

Despite the advances represented by works documenting the combination of generalized know-how and generalized know-why in operationally validated theories, much remains to be done in bringing this work into widespread practical use. Latour (1987) shows how metrological centers of calculation coordinate the calibration of reference standards and enroll various audiences as members participating in networked communities of inquiry and practice. Others (Daston, 1992; Shapin, 1994; Biagioli 1996) show how vital trust is in the history of science, since the credibility of agreement between theory and experiment depended largely on a moral economy and the effectiveness of unstated social norms. It is in fact difficult to imagine how any standard of communication or political association, whether a spoken language, an alphabet, radio transmission specifications, a profession, or the metric system, could be formed in the absence of what is broadly considered social capital (Coleman, 1988; Putnam, 2002).

These issues define the task before us. The planes are invented and proven, and aerodynamic theory is understood, but there are no airports, standardized aviation procedures, air traffic controllers or communications, or even sufficient taxis, hotels, and restaurants available to service those ready, able, and willing to take advantage of the new technology. The technical know-how and know-why cannot in and of themselves im-

prove the practice of measurement. For that, we need social capital. The availability of advanced training programs in, and introductory books on, Rasch measurement; widespread demands for increased accountability in education, health care, government, and corporate practice; research on the philosophy and history of effective measurement; the comparability of results across repeated applications of the same models to different instruments measuring the same thing; publishers' products integrating Rasch-calibrated assessments with instructional and clinical applications; and higher expectations as to what can be achieved in measurement are all contributing to increases in Rasch measurement's social capital.

It would seem that we are in the process of making a transition from a stage dominated by early adopters to one dominated by a broader audience of new, younger measurement consumers, and more conservative users, just as the 1990s was marked by the transition from a stage dominated by the original innovators to that of the early adopters. This conjecture is supported by the fact that Bond and Fox's (2001) highly accessible Rasch measurement text sold out at the American Educational Research Association's annual meeting again in 2004, for the third year in a row.

Are Rasch measurement models just a handy set of statistical tools that provide some convenient mathematical properties when they happen to work? Many, perhaps most, of those using the models might agree. Others see in the models the potential for a fundamental clarification of the objects of the human sciences. After all, it is probably no coincidence that the second scientific revolution occurring in the nineteenth century (Brush, 1998; Kuhn, 1961/1977, pp. 219-20) followed closely on the heels of the emergence of the metric system and its consummation of the union of mathematics and measurement (Roche, 1998, p. 145; Fisher, 2003a).

Might a similar scientific revolution be in store for the human sciences in the wake of common metrics for reading, writing, and mathematics in education, for health and quality of life in health care, and for skills and innovation in business? What kind of a global political economy

might be possible if we can make the growth of human, social, and natural capital just as scientifically accountable as financial and manufactured capital is now? If health, happiness and joy, fun and loving kindness, truly re-creative entertainment, and productive functionality are really as measurable as they seem to be in so many Rasch applications, and the social capital needed for mobilizing these measures is made available, we might yet see a new epoch in the history of humanity, one marked by a more consistent integration of theory and practice, of mathematics and philosophy, of peace, justice, and spiritual satisfaction. The greatness of a people is measured by the dimensions of its aspirations. Dare we aspire to be the measure of all things? Dare we not?

References

- Biagioli, M. (1996). Etiquette, interdependence, and sociability in seventeenth-century science. *Critical Inquiry*, 22(2), 193-218.
- Bond, T., and Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates [<http://homes.jcu.edu.au/~edtgbook/>].
- Brush, S. G. (1998). *The history of modern science: A guide to the second scientific revolution, 1800-1950*. Ames, IA: Iowa State University Press.
- Burdick, H., and Stenner, A. J. (1996). Theoretical prediction of test items. *Rasch Measurement Transactions*, 10(1), 475 [<http://www.rasch.org/rmt/rmt101b.htm>].
- Carpenter, P. A., Just, M. A., and Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97, 404-431.
- Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94 (Supplement), 95-120.
- Daston, L. (1992). Baconian facts, academic civility, and the prehistory of objectivity. *Annals of Scholarship*, 8, 337-363. Report in L. Daston,

- (Ed.) (1994). *Rethinking objectivity* (pp. 37-64). Durham, NC: Duke University Press.
- Dawkins, R. (1976). *The selfish gene*. Oxford, England: Oxford University Press.
- Dawson, T. L. (2002, Summer). A comparison of three developmental stage scoring systems. *Journal of Applied Measurement*, 3(2), 146-89.
- Dawson, T. L. (2002, March). New tools, new insights: Kohlberg's moral reasoning stages revisited. *International Journal of Behavioral Development*, 26(2), 154-66.
- Embretson, S. E. (1998, September). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3), 380-396.
- Fisher, W. P., Jr. (2003a). The mathematical metaphysics of measurement and metrology: Towards meaningful quantification in the human sciences. In A. Morales (Ed.), *Renascent pragmatism: Studies in law and social science* (pp. 118-53). Brookfield, VT: Ashgate Publishing Co.
- Fisher, W. P., Jr. (2003b, December). Mathematics, measurement, metaphor, metaphysics: Part I. Implications for method in postmodern science. *Theory and Psychology*, 13(6), 753-90.
- Fisher, W. P., Jr. (2003c, December). Mathematics, measurement, metaphor, metaphysics: Part II. Accounting for Galileo's "fateful omission." *Theory and Psychology*, 13(6), 791-828.
- Gladwell, M. (2000). *The tipping point: How little things can make a big difference*. Boston: Little, Brown, and Company.
- Green, K. E., and Kluever, R. C. (1992). Components of item difficulty of Raven's Matrices. *Journal of General Psychology*, 119, 189-199.
- Green, K. E., Kluever, R. C., and Wright, B. D. (1994). Predicting item difficulties from item characteristics. *Rasch Measurement Transactions*, 8(2), 354 [http://www.rasch.org/rmt/rmt82c.htm].
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
- Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis*, 52(168), 161-193. Report in T. S. Kuhn, (Ed.) (1977). *The essential tension: Selected studies in scientific tradition and change* (pp. 178-224). Chicago: University of Chicago Press.
- Lapr e, M. A., and Van Wassenhove, L. N. (2002, October). Learning across lines: The secret to more efficient factories. *Harvard Business Review*, 80(10), 107-11.
- Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. New York: Cambridge University Press.
- Latour, B. (1995). *Cogito ergo sumus!* Or psychology swept inside out by the fresh air of the upper deck: Review of Hutchins' *Cognition in the Wild*, MIT Press, 1995. *Mind, Culture, and Activity: An International Journal*, 3(1), 54-63.
- Law, J. (2002). *Aircraft stories: Decentering the object in technoscience*. Durham, NC: Duke University Press.
- Lewin, K. (1951). *Field theory in social science; selected theoretical papers* (D. Cartwright, Ed.). New York: Harper and Row.
- Markward, N. (2004). Establishing mathematical laws of genomic variation. *Journal of Applied Measurement*, 5(1), 1-14.
- Markward, N., and Fisher, W. P., Jr. (2004). Calibrating the genome. *Journal of Applied Measurement*, 5(2), 129-41.
- Putnam, R. D. (2002). *Democracies in flux: The evolution of social capital in contemporary society*. New York: Oxford University Press.
- Roche, J. (1998). *The mathematics of measurement: A critical history*. London: The Athlone Press.
- Shapin, S. (1994). *A social history of truth: Civility and science in seventeenth-century England*. Chicago: University of Chicago Press.
- Stenner, A. J., and Stone, M. (2003). Item specification vs. item banking. *Rasch Measurement Transactions*, 17(3), 929-30.

Wegner, D. (1991). Transactive memory in close relationships. *Journal of Personality and Social Psychology*, 61(6), 923-9.

Wright, B. D. (1994, Summer). Theory construction from empirical observations. *Rasch Measurement Transactions*, 8(2), 362 [<http://www.rasch.org/rmt/rmt82h.htm>].

Wright, B. D., Stenner, A. J., and Vanezky, R. (1995, Winter). Reading in America: Stenner's Lexiles confirmed. *Rasch Measurement Transactions*, 8(4), 387-388 [<http://www.rasch.org/rmt/rmt84a.htm>].

Wright, B. D., and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.