 DOI: 10.1051/metrology/201712004

Towards an alignment of engineering and psychometric approaches to uncertainty in measurement: Consequences for the future

William P. Fisher, Jr.[1,*], *A. Jackson* Stenner[2]

[1]LivingCapitalMetrics.com, Sausalito, California 94965 USA & BEAR Center, University of California, Berkeley, California 94720 USA

[2]MetaMetrics, Inc., Durham, North Carolina 27713 USA

Abstract. The International Vocabulary of Measurement (VIM) and the Guide to Uncertainty in Measurement (GUM) shift the terms and concepts of measurement information quality away from an Error Approach toward a model-based Uncertainty Approach. An analogous shift has taken place in psychometrics with the decreasing use of True Score Theory and increasing attention to probabilistic models for unidimensional measurement. These corresponding shifts emerge from shared roots in cognitive processes common across the sciences and they point toward new opportunities for an art and science of living complex adaptive systems. The psychology of model-based reasoning sets the stage for not just a new consensus on measurement and uncertainty, and not just for a new valuation of the scientific status of psychology and the social sciences, but for an appreciation of how to harness the energy of self-organizing processes in ways that harmonize human relationships.

1 Introduction

While only rarely ever thinking about it, we use language to manage the uncertainty of the future. Everyone benefits from being able to connect words, ideas, and things in the world—to communicate—but without knowing very much about how the language in use came to be, or why it is written, pronounced, or structured as it is. We have simply learned that we can rely on our words to mean about the same thing the next time we use them as they did the last time. We derive a great deal of security from knowing we can manage the uncertain future using words that have served us well in the past. If memory failed us all, or if it was impossible to link concepts and things with some material visual or auditory representation, we would have to invent new words for things in a constant process of re-invention. In that kind of world, life's unpredictability would make experience very different from what it has been for humanity.

Science extends and refines language in ways enabling the management of new, previously inaccessible tasks. Theoretical conceptualizations and experimental

* Corresponding author: william@livingcapitalmetrics.com or wfisher@berkeley.edu

Article available at http://cfmetrologie.edpsciences.org or https://doi.org/10.1051/metrology/201712004

substantiations of new phenomena, such as disease-causing germs and electricity, are embodied in technologies like vaccines and appliances distributed throughout interconnected networks [1,2]. Science systematically associates new words with new concepts in order to bring new things into the social world [3,4]. These word-concept-thing assemblages are kept in close contact with standards by technicians trained in the creation and use of the relevant tools. But all of this effort is expended so that end users can employ the new words and ideas in the same way as any other words, which is to say, without understanding the technical details of exactly how specifically unforeseeable future connections with something real in the world were made predictable and manageable.

Though it is not often articulated in this way, uncertainty and its management are plainly a matter of central importance in science. Measurement is the crucial activity that brings together in a portable technology (an instrument) what was learned in the past from data. That learning hinges, first, on the data being explained well enough by theory to predict future observations, and second, on the reduction in uncertainty realized in the precision of the instrument calibration. The calibration of new classes of instruments intended to measure previously unknown phenomena is, then, fundamentally a process of bringing new things into language by establishing consistent and reproducible relationships between their properties and the theoretical ideas and words instrumental to their communication. Science goes beyond everyday language in ascertaining and systematically reducing the uncertainty of the number words representing quantities, and in so doing opens up new opportunities for the creation of shared meanings and communities.

2 Uncertainty in metrology and psychometrics

It is in this context that we come to the most recent editions of the International Vocabulary of Metrology (VIM) and the Guide to Uncertainty in Measurement (GUM), which document a shift in the terms and concepts used in communications on uncertainty and measurement information quality [5,6]. The change is away from an Error Approach (also known as the Traditional Approach or the True Value Approach) and in favour of an Uncertainty Approach. Instead of assuming that the goal of measurement is the closest possible estimation of an unknown true value (the Error Approach), metrology now holds that measurement information supports only an assignment of a range of values, given that no mistakes have been made. This range varies depending on what information is taken into account, leading to the development of uncertainty budgets [7,8].

An analogous shift has taken place in psychometrics, where recent research presented at a series of symposia jointly sponsored by several International Measurement Confederation (IMEKO) technical committees (TC-1, TC-7, and TC-13) suggests provocative new practical and theoretical correspondences between metrology and psychometrics [9-12]. In accord with those similarities, differences between the psychometric binomial model's True Score Theory (and associated Classical Test Theory) and measurement theoretical approaches to error/uncertainty [9] parallel aspects of the shift documented in the VIM and GUM.

First, both metrology's Error Approach and psychometrics' True Score Theory focus on an assumed distinction between random and systematic errors. In both metrology and psychometrics, these sources of error are always assumed distinguishable, should be treated differently, and cannot be combined to form a total error. Second, both metrology's Uncertainty Approach and developments in psychometrics involving the evaluation of uncertainty relative to interval units of measurement focus on mathematical treatments of measurement uncertainty by employing an explicit measurement model characterizing the measurand in terms of an essentially unique value.

 DOI: 10.1051/metrology/201712004

The GUM and IEC documents provide guidance on the Uncertainty Approach to end users as to the case of a single reading of a calibrated instrument. No such guidance has yet been systematically available in psychometrics, in large part because instruments calibrated and traceable to uniform and universally available consensus unit standards are still unusual, though not unknown [10]. Routine estimation of individual measurand uncertainties is, then, encumbered by widespread reliance on True Score Theory's scale- and sample-dependent ordinal score units, and associated statistical methods. True Score Theory's sampling approach to group-level uncertainty supports one motivation for the probabilistic form of statistical models evaluated via significance tests, while measurement theory is instead motivated by the response process itself in deriving a scientific model evaluated in terms of explanatory theory, meaningfulness, ethical criteria, and practical utility [13-16]. The latter is then able to characterize unit traceability in terms of a unique value systematically qualified relative to the effects of local uncertainty, model fit, bias and DIF, range restriction, etc.

3 The contrasting shapes of uncertainty

One especially salient contrast between True Score Theory's and psychometric measurement theory's approaches to uncertainty concerns their different U-shaped vs arch shaped error distributions (Fig. 1) [17]. This contrast stems from the lack of expectations concerning the response structure in True Score Theory, and the explicit expectations for it

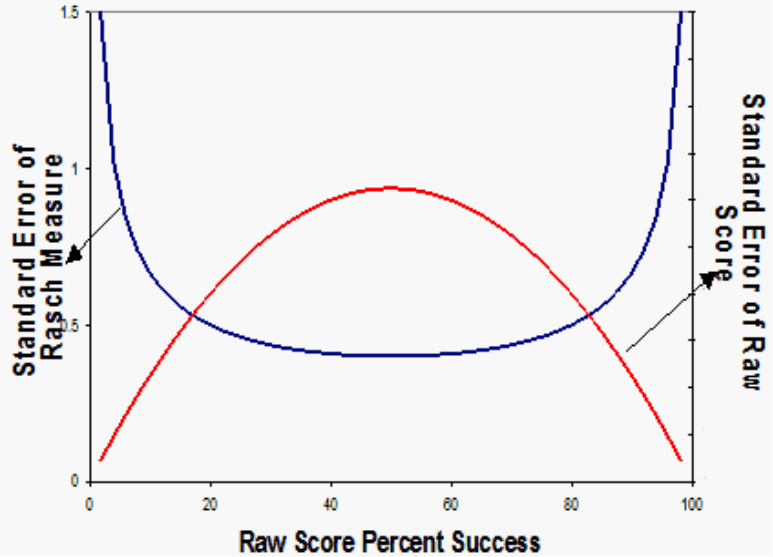

Fig. 1. Standard errors of measures vs scores [17].

in measurement theory [18]. In True Score Theory, the motivation for making models probabilistic is rooted in statistical sampling. Items are assumed to be of equal difficulty. The normal distribution is assumed, meaning that the tendency toward the mean dominates model conceptualization. Uncertainty is contingent on the probability of a response being near the middle part of the distribution, where there are 50-50 odds of success, and not toward one extreme or the other, where scores of 0% or 100% have the lowest possible uncertainty.

In measurement, in contrast, variation in item difficulties is expected and modelled, with the goal of obtaining a sufficient statistic, one "that summarizes the whole of the relevant information," as Ronald Fisher put it in 1922 [19]. Statistical sufficiency formulated at the individual level, per Rasch [20,21], is conceptually identical with invariance in measurement [22,23].

The identification of invariant profiles in the difficulty or agreeability of test and survey items enables use of the item location hierarchy in qualitative, substantive interpretation and theory-building otherwise inaccessible to psychological and social measurement [24-27]. A close match between a student's measure and a test item's difficulty, for instance, indicates the border between the instructional materials the student has mastered (the easier items on which the student has a high probability of success), and those not yet mastered (the harder items on which the student has a low probability of success).

Uncertainty in this context becomes useful in allocating resources appropriate to the budgeting constraints of varying applications. Instructionally useful formative applications in the classroom, for instance, may be diagnostically valid for day to day use within wider confidence limits than can be tolerated for quality improvement, research, or accountability applications. Fewer test items and more uncertainty can be tolerated in individual applications given the background knowledge and familiarity with the student possessed by the teacher. High stakes assessments in which new interventions are being evaluated, or that are used as the basis for graduation, admissions, or other decisions, must necessarily obtain more and higher quality information with less uncertainty, typically by asking more questions in a more rigorously controlled environment.

The statistical sampling approach to uncertainty taken by True Score Theory, in contrast, is less applicable to instructional problems. Uncertainty is lowest when scores are at the extremes, when the student knows nothing or everything, but this minimal uncertainty does nothing to indicate either how much too difficult the test items are, or how much too easy they are.

4 The role of uncertainty in unit definition

Andrich [9] shows the role played by randomly unimodal error distributions with smooth density functions in the formulation and realization of measurement invariance using Rasch's probabilistic models. Ordered categories for scoring responses to test and survey items are required to partition a range of variation meaningfully to be useful as a basis for measurement. True Score Theory stops with these ordinal categories and scores, manipulating the numbers as though they stand for a qualitatively meaningful and substantive property (intelligence, achievement, health, etc.) that adds up like the numbers do. Rasch models interval properties in a way useful for experimentally testing ordinal data to assess its capacity for substantiating the hypothesis of an additive, invariant unit. Asserting the need for such a model, and for experimental tests of the existence of a unit of comparison, goes against the grain of the methods typically employed in psychology and the social sciences. Considerable resistance to the use of Rasch's models has come from those wedded to simpler and less demanding numerical methods, though the scientific defensibility and practical utility of Rasch-based results has led to the routine use of the models in many high-stakes domains of research and practice.

Broader scale applications of Rasch's models are likely to follow from the results described by Andrich [9]. What he proposes should decisively resolve the "attenuation paradox" [28,29], namely, that higher reliability coefficients are not necessarily associated with better measurement. In True Score Theory, a perfectly discriminating test, for instance, is one that separates all examinees into two groups, one with no correct responses and the other with all correct responses. This can be seen to follow from True Score

Theory's peaked uncertainty distribution, where uncertainty is lowest at the extremes (Fig. 1).

The situation becomes more complex in the context of Guttman's deterministic approach to measurement. Here, the expectation of an ordered sequence of item difficulties contrasts with True Score Theory's supposition of equal item difficulties. But after an incorrect response is obtained, Guttman expects there to be no further correct responses to more difficult items. This rarely occurs in practice, so large percentages of data are often deemed unscalable in Guttman applications, leading to the method's rare application.

Rasch's models are basically a probabilistic formulation of Guttman's requirement of monotonic consistency in item responses [30,31]. The attenuation paradox takes a more specific form in the Guttman context than it does in True Score Theory. Guttman requires perfect reliability and discrimination in every score group, so that everyone with a score of 3 correct responses to 10 questions has exactly the same pattern. The end result, however, is that it becomes impossible to formulate any method for estimating how much more difficult one item is than another. This problem follows from the lack of information, the absence of any kind of a stochastic resonance, from which the magnitude of the difference might be estimated.

Duncan [32] accordingly observes that "It is curious that the stochastic model of Rasch, which might be said to involve weaker assumptions than Guttman uses [in his deterministic models], actually leads to a stronger measurement model." Andrich's [9] explanation of the role randomly unimodal error distributions play in applications of Rasch's models may satisfy Duncan's curiosity. It seems to me that Andrich's account is likely to turn out to be an image of the deep structure of how certain kinds of noise-induced order [33, 34] come about. To what extent are the "sufficient conditions for a system to exhibit stochastic resonance" [35] the same as or different from the conditions sufficient for fit to a stochastic Rasch measurement model [15,16,20,21,23]? How similar or different from the True Score Theory-Rasch or Guttman-Rasch contrasts is "the transition from chaotic behavior to ordered behavior induced by external noise…observed in a certain class of one-dimensional mappings" [33]? To what extent could individual-based ecological models and agent-based economic models [36] be informed by Rasch's individual-level stochastic measurement models based in sufficient statistics?

This phenomenon of stochastic resonance provides a metaphor that meets the terms of Galison's [37] search for a way to talk about the unity-through-disunity observed in the social and conceptual discontinuities characterizing some complex systems, such as science [38-41]. Galison notes the insufficiency of Peirce's and Wittgenstein's multistrand cable and thread analogies as images of how communities of theoreticians, experimentalists, and instrumentarians interact. He suggests that instead of these homogenous images in which the whole is the sum of the parts, we need images of discontinuous structures, like amorphous semiconductors with disordered atomic properties, or laminated structural engineering materials. In these cases, microscopic failure and disorder provide otherwise-unattainable signal-noise ratios and structural integrity. A great deal more research and study is needed to see if the resonance properties exhibited in Rasch's stochastic measurement models' randomly unimodal error distributions [9] provide the combined harmonies and discordances needed to coordinate the varied material and symbolic processes creating the binding culture of science [37-41].

5 Uncertainty budgets

Andrich [9] notes that

Measurement in the social sciences has not reached a level where the degree of precision is routinely stated in advance. To be able to do so, substantive research in the

 DOI: 10.1051/metrology/201712004

construction of ordered categories will require the same detailed empirical research that natural scientists carry out in constructing their instruments. The PRM [Polytomous Rasch Model] provides a basis for assessing the precision of measurement achieved.

Theory-informed empirical research on reading ability has begun to approximate the level of detail obtained in the construction of instruments in the natural sciences [9-12,20,27,42]. Implications for uncertainty budgets and the practical interpretation of psychometric measures begin from a distinction between Type A and Type B uncertainties [5,8].

Type A uncertainties are statistical and random, whereas Type B uncertainties are usually systematic. The True Value Approach did not allow uncertainties to be combined, but the more recent Uncertainty Approach allows the estimation of a total uncertainty from the accumulation of Type A and Type B uncertainties. Uncertainty components are no longer classified as random or systematic because of this capacity to bring all uncertainty estimates together into a single frame of reference in which they are comparable.

What might an uncertainty budget for psychometric measures look like? Table 2 proposes some elements, following the model provided by Bucher [7]. Type A uncertainties may emerge primarily from sampling considerations. As to Type B, the usual modelled measurement uncertainty, estimated as a function of the number of items administered, is complemented by a version inflated by positive values of the mean square model fit statistic (a chi-square divided by the degrees of freedom) [43]. Additional sources of uncertainty introduced by range restriction or bias might be imagined.

Table 2. A possible psychometric uncertainty budget [modelled from example in 7]: How to fill in the blanks?

Type A Uncertainty						
	Uncertainty description	**Uncertainty**	**Distribution**	**Divisor**	**Standard uncertainty**	**Variance**
1	Repeatability					
2						
	Combined Type A Uncertainty					
	Type B Uncertainty					
	Uncertainty description	**Uncertainty**	**Distribution**	**Divisor**	**Standard uncertainty**	**Variance**
1	Modeled	1 / sqrt(√∑(Pni(1 -Pni))	Randomly unimodal			
2	Fit-inflated	Modeled * MnSq>1				
3	Range restriction					
4	Bias					
	Combined Type B Uncertainty					

Before uncertainty contributors can be combined (using a root sum square method), they must be random, independent, and normalized to a standard uncertainty using the divisor suited to the relevant distribution, as specified in the GUM [5].

6 Implications for a new art and science of self-organizing complex adaptive systems

Chaitin [44] observes that the deterministic conceptualizations of Newtonian mechanics and the arithmetic of the natural numbers have given way to chaos and randomness. Reductionist approaches can no longer tenably posit that information on individuals will aggregate into reliable information on groups. The whole is no longer the sum of the parts, even in the areas of life where that seems most obviously true [45]. Instead of fully unique, independent, and separable individual elements, even in physical phenomena like the combined gas law, structural presuppositions at the microlevel are unavoidable.

It seems as though a transition is occurring in fundamental concepts, a transition analogous to the difference between True Score Theory and Guttman, on the one hand, and Rasch models, on the other. True Score Theory and Guttman assume all individuals are interchangeable, that the whole is the sum of the parts, and that there is no need for structural presuppositions. Rasch, in contrast, shows that individuals are not entirely unique, that they exhibit interdependencies, that the whole is more than the sum of the parts, and that group-level structure consistently emerges whether or not one is looking for it. The question is how to set up media for the self-organized, bottom-up repeatable and reproducible display and communication of these complex adaptive multilevel structures [38,46]. New opportunities are opening up in this direction across the sciences. For instance, in a recent book on complexity economics, Arthur [46] points out that

> Science and mathematics are shedding their certainties and embracing openness…there is no reason to expect that economics will differ in this regard.

New agent-based models re-orient psychology and social science away from policies imposing mechanically aggregated statistical results from the top down toward bottom-up initiations of grassroots efforts. The overall effect is akin to the difference between quality control and continuous quality improvement methods in industry [47]. The former were characterized by "tail-chopping" methods that merely removed the undesired end of a quality distribution, just to have it filled again in the next iteration. The latter, "curve-shifting" methods empower everyone on the front line to act to improve the production system itself via bottom-up, organic, self-organizing approaches to creating genuine value [38,46]. Providing teachers, clinicians, and managers with the coherent information tools they need to manage their responsibilities across the discontinuities inherent to multilevel communications systems [47-49] will likely result in quality revolutions in education, health care, human resource management, and other areas equal to or greater than that experienced in manufacturing [50-52]. Realizing the practical value of measurements as objective bases for decision supports demands close attention to uncertainty [53-54]. Communicating the value of educational, health, human resource, and other psychological and social measures in common units with known uncertainties coordinated across levels of complexity is essential to grasping the opportunities for improved outcomes we have within our reach.

References

1. B. Latour, *The Pasteurization of France* (Harvard University Press, Cambridge, Massachusetts, 1988)
2. B. Latour, *Reassembling the social* (Oxford University Press, Oxford, England, 2005)
3. E. Hutchins, Philosophical Psychology **27**, 34 (2014)
4. N. Nersessian, Mind, Culture, and Activity **19**, 222 (2012)

5. Joint Committee for Guides in Metrology (JCGM/WG 1) *Evaluation of measurement data--Guide to the expression of uncertainty in measurement* (International Bureau of Weights and Measures—BIPM, Sevres, France, 2008)
6. Joint Committee for Guides in Metrology (JCGM/WG 2) *International vocabulary of metrology: basic and general concepts and associated terms, 3rd ed (with minor corrections)* (International Bureau of Weights and Measures—BIPM, Sevres, France, 2012)
7. J. L. Bucher, *The metrology handbook* (ASQ Quality Press, Milwaukee, Wisconsin, 2012)
8. C. Ratcliffe, B. Ratcliffe, pp. 33-37 in *Doubt-free uncertainty in measurement* (Springer International Publishing, Zurich, Switzerland, 2015)
9. D. Andrich, J Phys.: Conf. Series in press (2017)
10. W. P. Fisher, Jr., A. J. Stenner, Measurement **92**, 489 (2016)
11. L. Mari, M. Wilson, Measurement **51**, 315 (2014)
12. L. Pendrill, W. P. Fisher, Jr., Measurement **71**, 46 (2015)
13. O. D. Duncan, Contemp. Sociol. **21**, 667 (1992)
14. O. D. Duncan, M. Stenbeck, pp. 1-35 in C. C. Clogg (Ed.), *Sociological Methodology 1988* (American Sociological Association, Washington, DC, 1988)
15. M. R. Wilson, Psychometrika **78**, 211 (2013)
16. D. Andrich, pp. 7-16 in J. A. Keats, R. Taft, R. A. Heath & S. H. Lovibond (Eds.), *Mathematical and Theoretical Systems* (Elsevier Science Publishers, North Holland, 1989)
17. J. M. Linacre, Rasch Meas. Trans. **20**, 1086 (2007)
18. M. R. Wilson, Measurement **46**, 3766 (2013)
19. R. A. Fisher, Philos. Trans. Royal Soc. London A **222**, 309 (1922)
20. B. D. Wright, Educ. Meas. Issues and Pract. **16**, 33 (1997)
21. G. Rasch, *Probabilistic models for some intelligence and attainment tests* (Danmarks Paedogogiske Institut, Copenhagen, 1960)
22. W. J. Hall, R. A. Wijsman, J. K. Ghosh, Annals Math. Statistics **36**, 575 (1965)
23. E. B. Andersen, Psychometrika **42**, 69 (1977)
24. B. D. Wright, M. H. Stone, *Best test design* (MESA Press, Chicago, 1979)
25. B. D. Wright, G. N. Masters, *Rating scale analysis* (MESA Press, Chicago, 1982)
26. M. Wilson, *Constructing measures* (Lawrence Erlbaum, Mahwah, New Jersey, 2005)
27. A. J. Stenner, W. P. Fisher, Jr., M. H. Stone, D. S. Burdick, Frontiers in Psychol.: Quantit. Psychol. Meas. **4**, 1 (2013)
28. J. Loevinger, Psychological Bulletin **51**, 493 (1954)
29. G. Engelhard, Rasch Meas. Trans. **6**, 257 (1993)
30. D. Andrich, Educ. Res. Persp. **9**, 95 (1982)
31. D. Andrich, p. 33-80 in N. B. Tuma (Ed.), *Sociological methodology 1985* (Jossey-Bass, San Francisco, 1985)
32. O. D. Duncan, *Notes on social measurement* (Russell Sage Foundation, New York, 1984)
33. K. Matsumoto, I. Tsuda, I. J. Statistical Physics **31**, 87 (1983)

34. L. Gammaitoni, P. Hanggi, P. Jung, F. Marchesoni, Reviews of Modern Physics **70**, 223 (1998)
35. S. M. Hess, A. M. Albano, Internat. J. Bifurcat. Chaos **8**, 395 (1998)
36. V. Grimm, S. F. Railsback, *Individual-based modeling and ecology* (Princeton University Press, Princeton, New Jersey, 2013)
37. P. Galison, *Image and logic* (University of Chicago Press, Chicago, 1997)
38. W. P. Fisher Jr, Procedia Computer Science in press (2017)
39. W. P. Fisher Jr, M Wilson, Pensamiento Educativo **52**, 55 (2015)
40. W. P. Fisher Jr, Measurement: Interdisc. Res. & Persp. **9**, 46 (2011)
41. W. P. Fisher Jr, Rasch Meas. Trans. **5**, 186 (1992)
42. L. Pendrill, NCSLi Meas.: The J. Meas. Sci. **9**, 22 (2014)
43. B. D. Wright, Rasch Meas. Trans. **9**, 436 (1995)
44. G. J. Chaitin, Internat. J. Bifurcat. Chaos **4**, 3 (1994)
45. A. Garfinkel, Reductionism pp. 443-459 in R. Boyd, P. Gasper, J. D. Trout (Eds.), *The philosophy of science* (pp. 443-459) (MIT Press, Cambridge, Massachusetts, 1991)
46. W. B. Arthur, *Complexity and the economy* (Oxford University Press, New York, 2014)
47. W. P. Fisher Jr, Assessment and Learning **2**, 6 (2013)
48. W. P. Fisher Jr., E. P.-T. Oon, S. Benson, J. Physics: Conf. Series in press (2017)
49. S. L. Star, K. Ruhleder, Info. Sys. Res. **7**, 111 (1996)
50. W. P. Fisher Jr., Standards Engineering **64**, 1 (2012)
51. W. P. Fisher Jr., Measurement **42**, 1278 (2009)
52. W. P. Fisher Jr., J. Appl. Meas. **12**, 49 (2011)
53. L. Pendrill, Metrologia **51**, S206 (2014)
54. W. P. Fisher Jr., B. Elbaum, W. A. Coulter, J Physics Conf. Series **238**, 012036 (2010)