William P. Fisher Jr. and Leslie R. Pendrill

1 Introduction: imagining the viability, feasibility, and desirability of extending the SI to include the psychological and social domains

Abstract: Metrology – quality-assured measurement – provides a commonsense way of connecting scientific and mathematical thinking with everyday thinking. Today's predominant approaches to quantification in psychology and the social sciences are inadequate to the urgent challenges humanity faces. Instead of hollow imitations of the measurement methods employed in physics, a metrological perspective grounds measurement across the sciences in cognitive and social processes familiar to all. Instead of foregrounding quantification as the primary determinant of measurement thinking, a more productive path forward orients social and psychological measurement toward the same sources in everyday thinking that were extended into science by physicists and engineers. In this introduction to *Models, Measurement, and Metrology Extending the SI: Trust and Quality Assured Knowledge Infrastructures*, we take up misconceptions about measurement, describe the chapters in this volume, summarize the history of developments in measurement theory, recount some recent social history, and generally pursue some implications of the metrological shift in perspective.

Keywords: psychological measurement, measurement modeling, metrological traceability, history of measurement, philosophy of science, psychometrics, Rasch models, quality assurance

1.1 To begin

As an introduction, our aim is to describe a metrological frame of reference for the other chapters in this book. A metrological context for psychological and social measurements will certainly sound, at the very least, counterintuitive to many readers. The very notion blatantly contradicts the analytical, statistical, and ordinal methods of quantification associated with achievement tests, surveys, rating scales, and assessments of various kinds.

Acknowledgment: We would like to thank the University of California, Berkeley Libraries and multiple chapter authors' independent funding sources for their support in making the book open access.

William P. Fisher Jr., BEAR Center, Berkeley School of Education, University of California, Berkeley; Living Capital Metrics LLC
Leslie R. Pendrill, Metrology, RISE Research Institutes of Sweden

https://doi.org/10.1515/9783111036496-001

Metrology – which has been formulated predominantly in the physical sciences – structures measurements read from instruments calibrated to quality-assured interval or ratio unit standards and disseminates traceability throughout networks of end users in support of interoperability. In psychology and the social sciences, in contrast, quantitative methods have traditionally focused on centrally planned and controlled ordinal data gathering and analysis. That shift in context of metrology to include psychological and social domains requires some elaboration, as many readers will not only be unfamiliar with it, they understandably may also find it inappropriate, misleading, or misguided.

This chapter briefly addresses some common misconceptions and unexamined preconceptions about measurement and metrology. This not only will draw out explications of longstanding and proven technical capabilities but will also highlight associated social and moral capabilities that must be included in any account of this kind. These latter are not rationalizations added on after the fact as ad hoc justifications for a predetermined purpose but are integral to the logic, aesthetics, and ethics of measurement. Measurement is, after all, deeply rooted in social contexts requiring a shared sense of fair dealing, such as trade and taxation. Although these norms usually remain in the background as unexamined assumptions, they nonetheless influence and shape the form of the agreements and distinctions we make (Alder, 2002, p. 2). In this respect, the standards we set and accept as guidelines for respectable or criminal behavior define what we value. This book takes up the problem of whether the standards informing our governance, market, educational, health care, and environmental institutions today adequately represent our cultural self-image, or if perhaps we could do better at living up to the ideals represented by the symbolic scales of justice.

These considerations point us toward a key insight, namely, that "the development from the spoken language . . . to symbols and pictograms . . . to what we now understand as written language is a perfect standardization process" (Weitzel, 2004, p. 11). Taking the development of written language as a model for an extended SI affords some key advantages. After documenting a detailed history of various efforts aimed at improving the human condition, Scott (1998, p. 357) recommends language as providing the best model for adaptable human institutions. He makes this recommendation because language is "a structure of meaning and continuity that is never still and ever open to the improvisations of its speakers." The fact that the existing SI undergoes continuous improvements while remaining in uninterrupted use speaks to its continued performance as an extension of everyday language.

To succeed, an extended SI must also possess exactly that kind of globally navigable comparability at the same time it accommodates local adaptations. The need for evolutionary potential in an extended SI will arguably be even more important than in the existing SI. This potential is realized by the demonstrated capacities of probabilistic models of measurement. Those models support empirical and theoretical estimations of measured quantities exhibiting persistent and reproducible invariances (within the range of uncertainty) across the shifting memberships of human populations and the contents of tests, item banks, and equated instruments. Measurement modeling in this context must encompass the complex demands of evolving circumstances, as the people involved and the challenges they face constantly change. It must also provide a context in which unique individuals can experience the powerful affirmation of feeling part of something larger than themselves without at the same time being overwhelmed by it. Given developments in psychological measurement theory and practice over the last 60 years and more, we feel confident this can be accomplished and has indeed been accomplished repeatedly over the last several decades.

This introduction cannot, of course, address the topic raised in anything approaching a comprehensive manner. Indeed, an entire book or series of books would be insufficient for that task. More fundamentally, one of the essential themes involved in this domain concerns the fact that the value of metrologically informed measurement is not, and cannot be, communicated primarily via merely cognitive forms of understanding. Lived experiences involving personal engagement, emotional associations, embodied sensations, economic and political consequences, creative opportunities for innovation, etc. will be required before the ideas offered here can be properly understood and evaluated.

Finally, though many positive benefits may be expected to accrue from new metrological infrastructures affording the mass distribution of higher quality and comparable forms of information, we do not regard the challenges involved as simple or the solutions as panaceas. We fully expect extending the SI into psychology and the social sciences to be controversial and difficult, and resource intensive. In this, we accept the lessons of history, recognizing that many of the original goals set over 200 years ago for the West's original unified system of decimalized metrics had to be abandoned or significantly modified. Those lessons also show that significant returns on investment were provided on massive up-scaling. The successes of those efforts have, however, altered the environment in which they occurred to such an extent that instead of enhanced prosperity they now threaten catastrophe. New ideas more explicitly cognizant of the organism-environment unity as the focus of natural selection are in order.

1.2 An audacious but grounded proposal: extending the SI

Given the usual understanding of quantitative methods as assigning numbers to operations, or more colloquially, reducing things to numbers, many readers will likely find the idea that the SI could be usefully, meaningfully, ethically, and productively extended to apply to human affairs, a preposterous, and even a dangerous, proposition. Given the typical perception that the natural world is measurable because it is inherently and automatically numerical, to many the notion of an extended SI applied to human subjectivity will sound positively ridiculous and not worth considering.

Readers with some background in the history and philosophy of science will have a somewhat – in some cases, a markedly – different perspective. This will especially be the case for those involved in science and technology studies, conceptual metaphor theory, cognitive and developmental psychology, and related areas. Even for these readers, however, the complexities of the issues may seem intimidating. Some few may find the idea that the SI could plausibly be extended into human affairs to invite curiosity.

Those readers familiar with technical developments in measurement theory and practice, finally, may be breathing a sigh of relief that the topic has finally been broached in a way that positions it as a matter for broad and serious consideration, and not as something to be dismissed out of hand. After all, the mathematical identity of measurement models applicable across the natural and social sciences was deemed "widely accepted" since the 1960s by two noted authorities almost 40 years ago (Narens & Luce, 1986).

One particularly lethal misconception about measurement must be immediately addressed: measure equations and quantity equations are not to be confounded; they are distinct concepts (Maxwell, 1873; de Courtenay, 2015). As can easily be shown, measurement is at the same time not fundamentally or primarily a matter of quantification (Mari et al., 2013, 2016). Many scientists, philosophers, and psychologists note that the inferential processes involved in mathematical thinking and quantitative modeling are already present in everyday language and thinking (Black, 1946, pp. 304–305; Bohr, 1963, p. 9; Einstein, 1954, p. 290; R. Fisher, 1935, p. 79; Guttman, 1994; Huxley, 1862, pp. 57–58; Nersessian, 2008). Metaphysically speaking, the intuition that the universe is somehow numeric in and of itself increasingly seems to be an incomplete and undeveloped initial insight into the idea that existence is broadly mathematical, and is so in profoundly logical, beautiful, and ethical ways. Understanding this does not require a high-level technical understanding or a convoluted esoteric philosophy. A commonsensical point of view is offered throughout this book, one that leads to end results devoid of hairsplitting logical arguments of limited applicability. Our arguments and those of the other authors of the chapters in this book instead open an array of highly pragmatic opportunities for the advancement of the arts and sciences and associated prospects for cultural and economic progress.

A hint that the SI could profitably be extended to include psychological and social domains might be found in the repeated observations of multiple measurement wheels reinvented *ad nauseam* in those areas over the last several decades. The reproduction of linearly comparable quantities across different studies employing different instruments on different samples was, in general, highly unexpected and not predicted by the researchers involved. But persistently surprising convergences keep coming to light. How are these results being produced? How can they be explained? Could they possibly be systematically developed into a general frame of reference for scientific methods? What might they mean for improved understanding and communications? What do they imply for social organization, and the relation of human beings to the Earth and the world of nature? To begin answering these questions, we should take a moment to retrace our steps and consider how we arrived at this point. We will focus in turn on issues of meaningfulness, economic incentives and rewards, and first philosophy, or metaphysics, after first giving an overview of the chapters in the book.

1.3 The chapters

The chapters in this book demonstrate how interval and ratio level measurements of social and psychological constructs are in principle capable of comprising an extended SI. This capability follows from the definition and estimation of unit quantities that retain their properties independent of particular samples and instruments. Quantitative measurements of this kind have been produced using well-established probabilistic models and methods for well over 50 years. The following chapters illustrate how to employ advanced measurement modeling to obtain one or more of several advantageous affordances.

Part I of this volume presents various chapters under the heading of Theory and Principles in Measurement and Metrology, while Part II takes up Designing and Calibrating Metrologically Viable Measurements: Methods and Applications. Leslie Pendrill (2024) sets the stage by following through from historical developments in psychology to present-day foundational studies of the applicability of traditional engineering of measurement systems and concept systems (quantities, units and relations between them) to human-based metrology. Matthew Barney and Feynman Barney (2024) then envision an integration of metrology, psychological measurement models, and artificial intelligence, describing how common metrics might be embedded in everyday life via quantities unobtrusively inferred from decisions and behaviors. Ernesto San Martin and colleagues (2024) explore the value of identified and partially identified models for inferring causal relationships in measurement. Jeanette Melin (2024) expands on the theme of causality in her investigation of validity as a matter of essential concern in an extended conceptualization of the SI. William P. Fisher, Jr. (2024) concludes Part I with reflections on some of the paradoxical reasons why metrological traceability to unit standards is not typically considered possible in social and psychological measurement, and how it is also eminently reasonable and urgently needed.

Part II's examples of metrologically viable measurement scaling begin with a contribution from Robert Massof, Chris Bradley, and Allison McCarthy (2024) showing how clinical signs and symptoms can be organized to model a continuous disease state variable. Steve Lang and Judy Wilkerson (2024) follow this with a description of their journey in the development of an affective assessment for measuring teacher dispositions, a domain not nearly as well investigated as that of cognitive or behavioral constructs. Harry Kollias (2024) contends with the often life-changing consequences of high-stakes assessments involving panels of judges who vary in their perceptions of task mastery and the satisfaction of performance standards. Greg Sampson and colleagues (2024) describe another high-stakes assessment context involving similar multifaceted problems in the licensure and certification of emergency medical technicians. Simon Karlsson and colleagues (2024) address municipal sustainability metrics with the overt intention of capitalizing on the metrological potentials of a distributed system of manageable metrics.

Trudy Mallinson (2024) asserts that "measurement is not a benign act" but has consequences for equity and social justice that demand close attention in devising and applying justice-oriented anti-racist criteria for validating quantitative results. Linda

Morell, Sean Tan, and Mark Wilson (2024) document the interrelationships of assessment content and inferences as to competencies in the measurement of twenty-first century skills. Dhanya Natha Kumar and Hrishi Joshi (2024) focus on surgical outcomes assessments and how they could better contribute to the advancement of patient-centered care if they were designed, calibrated, and maintained with closer attention to fundamental measurement principles. Finally, David Sul (2024) takes up the question as to how measurements need not elevate any single cultural worldview to a position from which it is allowed to negate or erase the values of any other worldview.

In different ways, all these chapters show how a meaningful extension of the SI to cover the psychological and social domains depends on and benefits from the following conceptual distinctions and methodological demands. Readers unfamiliar with technical issues in measurement and metrology should approach the chapters with the following in mind and might be motivated to explore these ideas at greater length if practical applications are to be undertaken:

- Ordinal scores are not interval measurements, just as numeric counts are not measured quantities (Wright, 1992b; Wright & Linacre, 1989). Everyone knows we cannot say who has more rock when I have two and you have five, yet we persist in fallaciously treating test scores as measurements in the absence of a defined unit quantity. Even though psychology and the social sciences have for several decades possessed the models and methods needed for defining and estimating quantities from counts, global institutions continue to assume counts and percentages of correct answers and of ratings suffice as quantitative measurements. This book argues that there are firm reasons based in theory and evidence for altering that circumstance.
- Stevens' (1946) initial categorization of four levels of measurement (nominal, ordinal, interval, and ratio) was later augmented by a fifth log-interval level (Stevens, 1957, 1959) recognized by Narens and Luce (1986) as in productive use across the natural and social sciences and in applications of additive conjoint measurement models like those described in this volume. In a recent review, Marmor and Bashkansky (2020) consider adding new types of quality data to the well-known nominal, ordinal, interval, and ratio scales, including for example ranked data, as previously proposed by Tukey (1986).
- Ordinal score data volume is reduced by interval measurement with no loss of information, as individual observations are both necessary and sufficient to the estimation of measurement model parameters (Andersen, 1977, 1999; Fischer, 1981).
- Physical measurements of length and distance, mass, and density expressed in SI units have been reproduced from ordinal observations via the application of probabilistic conjoint measurement models (Choi, 1998; Moulton, 1993; Pelton & Bunderson, 2003; Stephanou & Fisher, 2013).
- The reproducibility of patterns of concrete observations from abstract measurements and formal theory is never perfect, of course, so data displays revealing failures of invariance pertinent to end user interests in improved outcomes are of value in

bringing out qualitative exceptions to the rule (Allen & Pak, 2023; Chien et al., 2011, 2018; Fisher, Oon, & Benson, 2021; Wright & Stone, 1979; Wright et al., 1980).

- Substantive integrations of qualitative and quantitative data and methods in formative assessments delineate learning progressions and map variation to locate individuals or groups relative to where they have been, where they want to go, and what to do next (Black et al., 2011; Duckor & Holmberg, 2019; Morell et al., 2017).
- Repeatable and reproducible definitions of meaningful additivity are essential to models sufficient to the identification of constructs as independent, within the range of uncertainty, of the sample measured and the instrument used to measure (San Martin & Rollin, 2013; San Martin et al., 2024; Wright, 1997a/b).
- Despite persistent repetitions of the idea that Rasch's models for measurement are one-parameter IRT models, they have no connection whatsoever with item response theory (Wright, 1984, 1997b). As Cliff and Keats (2003, p. 15) recognized, it was only "by means of some highly dubious assumptions [that] the Rasch formulation was generalized to what is now called Item Response Theory."
  - Neither Rasch nor any of the major contributors to the development of measurement theory and practice based in Rasch's models assert a connection with IRT.
  - Despite the inclusion of a model equation of the same form as those described by Rasch, the meaning of the parameters is entirely different.
  - IRT rationalizes unidentified and internally contradictory item parameters in an overt prioritization of a modern, positivist focus on describing data presumed to exist independent of human interests.
  - This renders IRT constitutionally incompatible with measurement theory, as has long been asserted (Andrich, 1989; Cliff, 1992, p. 188; Embretson, 1996a; Lumsden, 1978, p. 22; San Martin et al., 2015; Verhelst & Glas, 1995, p. 235; Wood, 1978, p. 31).
- That the association of Rasch's models with IRT persists despite repeated explanations of the illogic involved may be another example of "effortless associative thinking" (Kahneman, 2003, 2011; Simon, 1997, 2000) characteristic of bounded rationality and exemplifying the powerful constraints imposed on thinking by the dominant paradigm:
  - Weiss (2021), for example, despite having "led the much of the seminal research behind Computerized Adaptive Testing, and trained several generations of eminent psychometricians," nonetheless
    - continues to categorize Rasch's models in IRT,
    - holds that Rasch's models have been "been replaced by more general [IRT] models that allow test items to vary in discrimination, guessing, and a fourth parameter,"
    - considers the models' invariance requirement to be an expendable assumption, and

    - concludes that Rasch's models "can best be viewed as an early historical footnote in the history of modern psychometrics."
  - This modernist IRT perspective is then inherently at odds with an unmodern metrological perspective, both in theory and in the facts of the continued rapid growth and development of Rasch-based theory and practice in the field (Aryadoust et al., 2019).
  - We may then have here an instance in which statistically oriented and measurement-oriented communities of research and practice will have to agree to disagree:
    - In the spirit of the "unified disunity" and "convergent divergence" (Blok et al., 2016, 2020; Bowker et al., 2015; Galison, 1997, 1999; Woolley & Fuchs, 2011) in thinking ineradicable from irreducible complexity, there is more to be gained from passionately held differences of opinion than from hollow and coerced consensus.
    - Of course, the unmodern paradigm does not eliminate metaphysics or bounded rationality from playing roles in science; it can only expand the limits in which "effortless associative thinking" obtains.
    - When it does, and new ecological economies of thought transform today's scientific, legal, market, and communications infrastructures, few will likely quibble over differences that do not make a difference.
- Multiple specialized software packages are available for testing empirical and theoretical invariance hypotheses and for estimating measurement and calibration parameters, for a wide variety of models (Adams et al., 2020; Andrich et al., 2017; Bulut, 2021; Doran et al., 2007; Hohensinn, 2018; Lamprianou, 2020; Li, 2006; Linacre, 2023b, 2024; Melin & Pendrill 2023; Pendrill 2024; Robitzsch et al., 2020; Torres Irribarra & Freund, 2014; Verhelst et al., 2007; Wilson et al., 2019).
- Comparisons of estimation algorithms and program outputs are available (Linacre, 2023a; Linacre et al., 2013; Robinson et al., 2019; Yumoto & Stone, 2011).
- The Rasch.org website offers the full texts of authoritative books, *Rasch Measurement Transactions*, and articles; information on software training workshops, and conference calendars.
- Rigorous separation of levels of complexity, where formal construct theory explains variation in abstract item calibrations and person measurements estimated from concrete observations, enable theoretical, metrological, and experimental issues to be dealt with by separate, collaborative communities of research and practice.
- Explanatory models that successfully predict and account for large proportions of variation in instrument calibrations and sample measurements make it possible to automate item generation on the fly and to infer ratings and measurements from observed behaviors (Barney & Barney, 2024; De Boeck & Wilson, 2004; Embretson, 2010; Fischer, 1973; Stenner et al., 2013) and in some cases can even enable metrological references for traceability (Pendrill, 2019, 2024; Melin et al., 2021).

– Finally, and perhaps most counterintuitively to many readers, culturally specific assessments can measure abilities and attitudes in terms meaningful to varied communities without compromising broader capacities for comparability (MacIntosh, 1998; Mallinson, 2024; Sul, 2024; Tennant et al., 2013; Teresi et al., 1995; Wilson, 1994a).

1.4 Some background history

1.4.1 Origins

L. L. Thurstone was a former electrical engineer who went into psychology, was the first president of the Psychometric Society, was a co-founder of *Psychometrika*, and was the director of examinations at the University of Chicago from 1924 to 1952. In 1928, he wrote:

> The scale must transcend the group measured.–One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid. A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement. (Thurstone, 1959, p. 228)

Introducing the conception and purpose of the measurement models he devised, on the first page of his 1960 book, Rasch expanded on this theme, saying:

> Individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments – tests or items or other stimuli – within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class – 'measuring the same thing' – independent of which particular individuals within a class considered were instrumental for the comparison. (Rasch, 1960, p. xx; also see Rasch, 1961, 1966a/b)

Rasch then intentionally formulated an individual-level model for measurement structured on the basis of observations of any kind of physical or psychological phenomenon:

> taken as nothing more than an accidental response, as it were, of an object – a person, a solid body, etc. – to a stimulus – a test, an item, a push, etc. – taking place in accordance with a potential distribution of responses – the qualification 'potential' referring to experimental situations which cannot possibly be [exactly] reproduced. (Rasch, 1960, p. 115)

Rasch showed that such response distributions “depended on one relevant parameter only,” one chosen so that the same multiplicative law applied no matter whether observations involved people, solid bodies, test items, or pushes. He concluded that:

> Where this law can be applied it provides a principle of measurement on a ratio scale of both stimulus parameters and object parameters, the conceptual status of which is comparable to that of measuring mass and force. Thus, . . . the reading accuracy of a child . . . can be measured with the same kind of objectivity as we may tell its weight. (Rasch, 1960, p. 115)

Four years later, in 1964, Luce and Tukey showed how:

> the fundamental character of measurement axiomatized in terms of concatenation [is extended to] qualitatively described ‘additivity’ over pairs of factors of responses or effects . . . [such that] the additivity is axiomatizable in terms of axioms that lead to scales of the highest repute: interval and ratio scales. (Luce & Tukey, 1964, p. 4)

They illustrated that measurement axiomatized on the basis of conjointly ordered pairs of factors “apply naturally to problems of classical physics and permit the measurement of conventional physical quantities on ratio scales,” concluding that:

> In the various fields, including the behavioral and biological sciences, where factors producing orderable effects and responses deserve both more useful and more fundamental measurement, the moral seems clear: when no natural concatenation operation exists, one should try to discover a way to measure factors and responses such that the ‘effects’ of different factors are additive. (Luce & Tukey, 1964, p. 4)

In 1986, Narens and Luce, then, generalized from examples spanning different physical and behavioral contexts (“the ordering by mass of objects characterized by their volume and density; the loudness ordering provided by a person for pairs of sounds, one to each ear; and the preference ordering provided by an animal for amounts of food at certain delays”). They showed that conjointly ordered effects of these kinds:

> not only provided a deep measurement analysis of the numerous nonextensive, ‘derived’ structures of physics, but also provided a measurement approach that appears to have applications in the nonphysical sciences and has laid to rest the claim that the only possible basis for measurement is extensive structures. (Narens & Luce, 1986, p. 177)

On the basis of these results, Narens and Luce (1986, p. 169) say that, with the introduction of the theory of additive conjoint measurement in the 1960s (Andersen, 1970; Brogden, 1977; Fisher & Wright, 1994; Green, 1986; Luce, 1959, 1978; Luce & Tukey, 1964; Newby et al., 2009; Pelton & Bunderson, 2003; Perline et al., 1977; Rasch, 1960, 1961, 1966a/b; Wright, 1968, 1977), the view that interval-scalable, fundamental measurement is possible for nonextensive structures became “widely accepted.” Andrich (1988, p. 22) concurred, pointing out that, “. . . when the key features of a statistical model relevant to the analysis of social science data are the same as those of the laws of physics, then those features are difficult to ignore.”

Despite the supposed wide acceptance of interval scalable results, these models remain far from defining mainstream measurement theory and practice. Wright (1968, 1977, 1997a/b, 1999; Wright & Masters, 1982; Wright & Stone, 1979, 1999) brought Rasch's ideas into fairly wide use, emphasizing that his models for measurement are not data models but are definitions of the laws of measurement (Wright, 1988). Wright and his colleagues and students (Bode et al., 2000; Chien et al., 2011; Connolly et al., 1971; Kielhofner et al., 2005; Linacre, 1997; Liu, 2018; Massof, 2008; Mead, 2009; Smith, 1994, 1997; Stenner et al., 2013; Wolfe et al., 2000; Wright, 1997a, 2012; Wright & Stenner, 1998; Wright & Stone, 1979) repeatedly addressed problems related to the definition of meaningful units and the design of instruments, applications, and reports incorporating them.

Wright (1997a, 2012; Wright et al., 1980; Wright & Stone, 1979) thought to leverage the inferential stability of established calibrations by making measurement and uncertainty estimates, along with graphical response consistency evaluations, available at the point of use as soon as observations were recorded. Given constructs proven via explanatory theory and experimental evidence to be stable across multiple samples and instruments, the next case through the door would not likely alter the definition of what was measured. Instead of assuming measurement is only ever a product of data analyses, repeated empirical validations of theoretical predictions afford the opportunity to devise self-scoring forms interpretable at the point of use. Wright may never have once included the word "metrology" in his writing, but he nonetheless articulated a fundamentally metrological goal when he saw how on-the-spot applications could be supported by measurements read from quality-assured instruments calibrated in a reference standard quantity.

In so doing, with no reference to the history of science, what Wright understood about measurement as the actionable modeling of the real world was well described by Ackermann (1985, pp. 143–144) when he noted that:

> Once clear statements of fact have been achieved through instrumental investigation, the reference of fact seems fixed and objective, and indeed it is. The world has been discovered to show a fixed and repeatable response in certain interactions as described in the language, and this response is an objective consequence of these interactions. . . . This process of achieving or constructing reference for language by development of a domain we will call the microprocessing of fact, after discussion of this phenomenon by Latour and Woolgar [1979]. When the process is complete, the evidence of microprocessing disappears, and mere correspondence, the very correspondence that has been slowly and carefully constructed, is all that remains.

Wright's admonitions and recommendations in this regard are not often followed, but the value of the probabilistic models of measurement he advocated was recognized by many soon after they were introduced (Wilson & Fisher, 2017). These models have been further explicated and increasingly applied in psychology, health care, and the social sciences over the last several decades (Andersen, 1977, 1980; Andrich, 1978, 2010; Andrich & Marais, 2019; Aryadoust et al.; 2019; Bezruczko, 2005; Boone & Staver, 2020; Embretson, 1996b, 2010; Engelhard, 2012; Fischer, 1973, 1981; Fischer & Molenaar, 1995; Fisher &

Wright, 1994; Hagell, 2014, 2019; Loevinger, 1965; Massof, 2008; Masters & Keeves, 1999; Pendrill, 2019, 2024; Melin et al., 2021; Pesudovs, 2006, 2010; Salzberger, 2009; Smith & Smith, 2004, 2007; von Davier & Carstensen, 2007; Wilson, 1992, 2018, 2023).

After expanding on developments in measurement theory dating back several centuries, including the introduction of Rasch's, and Luce and Tukey's, additive conjoint perspectives, Wright (1997b, p. 44) concludes his history of social science measurement saying that "Today there is no methodological reason why social science cannot become as stable, as reproducible, and hence as useful as physics." Although Wright (1997b, p. 33) recognized the social roots of uniform metrics in society's demands for fair taxation and trade, and though he clearly stated that "Science is impossible without an evolving network of stable measures," he did not reflect on the relevant metrological challenges or opportunities. O. D. Duncan, in contrast, "the most important quantitative sociologist in the world in the latter half of the twentieth century" (Goodman, 2007, p. 131), articulated a metrological connection with Rasch's models for measurement. Duncan then plays a key role in expanding on Wright's claim as to the methodological potential for social science to become as stable, reproducible, and useful as physics.

1.4.2 Otis Dudley Duncan's contributions

Over the course of the 1980s, Duncan introduced Rasch's measurement models into sociology. In so doing, Duncan (1984b, pp. 38–39) suggested "that social measurement should be brought within the scope of historical metrology, while that discipline learns to take advantage of sociological perspectives." Toward that end, Duncan argued that:

> What we need are not so much a repertoire of more flexible models for describing extant tests and scales . . . but scales built to have the measurement properties we must demand if we take 'measurement' seriously. As I see it, a measurement model worthy of the name must make explicit some conceptualization – at least a rudimentary one – of what goes on when an examinee solves test problems or a respondent answers opinion questions; and it must incorporate a rigorous argument about what it means to measure an ability or attitude with a collection of discrete and somewhat heterogenous items.
>
> Thurstone explicated the meaning of measurement as it might be accomplished by such an instrument. Rasch provided the formalization of that meaning. (Duncan, 1984b, p. 217)

Complementing that view on models implementing a rigorous conception of measurement, Duncan (1984b, pp. 206–207) gives a long list of "ambiguous and poorly discriminated concepts" as evidence of "the prevailing chaos in which there is a multiplicity of 'tests,' 'scales,' or 'instruments' ostensibly serving as 'measures'" but which fail to live up to even a generous sense of the word. He cites research showing:

> many instances of the same items (questions, or statements calling for an agree/disagree response) in tests intended to measure different constructs, different and dissimilar items in tests

> with the same or similar names, a widespread habit of arbitrarily modifying tests when applying them in new research (and thereby precluding comparison or any benefit of standardization), and the replacement of old scales by new ones without cross-calibration between them and without demonstration of improved validity. (Duncan, 1984b, p. 207)

The phenomenon of myriad incommensurable metrics was also once the case in physics, as Duncan brings out in his notes on historical metrology and has also long been amply demonstrated in the history of science (Alder, 2002; Ashworth, 2004; Black, 1962; Crosby, 1997; Hesse, 1970; Kula, 1986; Nersessian, 2008; Roche, 1998; Wise, 1995).

In this context, somewhat inadvertently, Duncan is here offering evidence that the dominant modern paradigm's conception of science as describing an independently given objective reality is fatally flawed. The question is, what methodological conclusions might we draw from the historical coevolution of metric standards with concurrent developments in conceptually aligned governance and economic principles, with the co-production of science and society (Bowker, 2016; Edelmann, 2022; Ihde, 1991; Jasanoff, 2004; Knorr Cetina, 1999; Power, 2004)?

The reductionist conception of quantity sees it as built up from elementary building blocks in the physical world. Here, wholes are the sums of parts. The associated idea of measurement as only describing a pre-existing given reality not only fails to hold in the history of the natural sciences but also undermines the very methodological foundations of psychology and the social sciences. To be sure, there is a marked and highly relevant distinction to be made between this naïve sense of an inherently quantitative universe and the objectively repeating and reproducible self-organized phenomena that persistently exhibit consistent properties across samples, instruments, laboratories, observers, time, and space. The point to be made is limited to noticing that uncritically held metaphysical faith in a transcendent reality falsely makes it appear that the incommensurability of metrics in sociology and psychology is a consequence of human subjectivity disconnected from objective reality. It is nothing of the sort.

Duncan's description of the chaos in social measurement indirectly amplifies the point made by Gödel (1931) and a wide range of others (Garfinkel, 1991; Lerner & Overton, 2017; Nagel & Newman, 1958; Toulmin, 1953; Wittgenstein, 1983) as to the existence of arithmetical truths that cannot be formally demonstrated. Gödel thought, and many others have agreed, that his proofs of this theorem ought to have been elevated to a fundamental principle of science equivalent to Einstein's theory of relativity (Calude, 2002, 2007; Chaitin, 1994; Floyd & Kanamori, 2016). In this vein, Holton points out that

> we can find even among the most hard-headed modern philosophers and scientists a tendency to admit the necessity and existence of a noncontingent dimension in scientific work. Thus Bertrand Russell speaks of cases 'where the premises of sciences turn out to be a set of presuppositions neither empirical nor logically necessary'; and in a remarkable passage, Karl R. Popper confesses very plainly to the impossibility of making a science out of only strictly verifiable and justifiable elements. (Holton, 1988, p. 41)

The idea that science can subject the totality of its presuppositions to experimental tests of truth or falsity has been roundly discredited for decades. Although no consensus has emerged as to what this means for a methodical logic of science (Gadamer, 1981, 1989; Nielsen & Lynch, 2022; Weinsheimer, 1985), there are certainly strong indications that an emphasis on the playful absorption into dialogical relationships offers a number of advantages for structuring a new perspective on the mutual implication of subject and object (Dawson et al., 2006; Fisher, 2004; Nersessian, 1996; Overton, 2002). Perhaps it is not unreasonable to begin systematically investigating other options offering alternatives to modern dualist assumptions of alienated subjects and objects (Fisher, 2019).

A great many cultural, economic, social, and psychological factors (Dewey, 1929; Faber, 1999a/b; Gigerenzer, 1993; Haraway, 2022; Kauffman & Roli, 2023; Kline, 1980; Overton, 2002; Prigogine, 1986, 1997; Prigogine & Stengers, 1984) contribute to the reasons why a new scientific paradigm has yet to cohere. But the time is past for clinging to counterproductive and obsolete ideas and methods, especially given the presence of viable alternatives. If quantitative methods merely accept naturally given units, then there would be no historical variation in the definitions of metric units or of the physical and social constructs worthy of investigation. Duncan then offers the relevant observation that:

> All measurement is . . . social measurement. Physical measures are made for social purposes and physical dimensions may be used by . . . scientists [in any domain of research or practice]. But social measurement in a narrower sense deals with phenomena that are beyond the ken of physics. To extend historical metrology to include social measurement, therefore, will require some modification of thought patterns. For one thing, we shall have to overcome our tendency to think of social measurement or quantification as something external to the social system in the sense, say, that the tailor's tape measure is external to the customer's waist. On the contrary, I argue, the quantification is implicit – sometimes explicit, for an observer not blinded by methodological preconceptions – in the social process itself before any social scientist intrudes. (Duncan, 1984b, pp. 35–36)

The roots of social measurement in social processes can be traced from the ancient Greek origins of mathematical thinking in Plato's accounts of Socratic dialogue (Fisher, 1988, 1992, 2003a/b, 2004, 2010). Rigorous conceptions of qualitatively meaningful quantification based in irreducible complexity (Commons et al., 2014; Dawson et al., 2006; Dawson-Tunik et al., 2005; Fischer & Dawson, 2002; Overton, 1998, 2002) can be seen to extend everyday language into mathematical scientific language (Fisher, 2019, 2020, 2021, 2023).

The social processes leading to the production of this volume involved a personal communication about Duncan between one of the authors (Fisher) and Benjamin Wright in the late 1980s. Wright stated that he had been unable to persuade Duncan to adopt Rasch's models during in their early years as young professors at the University of Chicago in the 1960s. Duncan later, however, grasped the essential differences between scientific and statistical models, worked through his understanding of the models from the bottom up, and forcefully advocated the adoption of Rasch's perspective in quantitative sociology (Duncan, 1984a/b/c, 1992; Duncan & Stenbeck, 1987, 1988).

But contrary to the arguments made by others as to the value of Rasch's ideas, Duncan did not frame his measurement perspective in terms of the choices made among models for data analyses. Instead, he was cognizant of the challenges and opportunities posed by metrological unit definitions:

> With the possible and, in any event, limited exception of economics, we have in social science no system of measurements that can be coherently described in terms of a small number of dimensions. Like physical scientists, we have thousands of 'instruments,' but these instruments purport to yield measurements of thousands of variables. That is, we have no system of units (much less standards for them) that, at least in principle, relates all of the variables to a common set of logically primitive qualities. There are no counterparts of mass, length and time in social science . . . . To the physical dimensions, economics adds money . . . . The fact that social science (beyond economics) does not have such a system of measurements is, perhaps, another way of saying that theory in our field is fragmentary and undeveloped, and that our knowledge is largely correlational rather than theoretical.

Significant advances in the development of predictive theories and explanatory models of measured constructs have occurred in the years before and since Duncan wrote (e.g., see Commons et al., 2014; De Boeck & Wilson, 2004; Embretson, 2010; Fischer, 1973; Green & Smith, 1987; Smith, 1996; Stenner & Smith, 1982; Stenner et al., 2013, 2016, 2023). Consistently reproducible correspondences of theory and evidence may be key factors substantiating a basis for confidence in systems of measurements traceable to a new class of candidate SI units. Documented instances (Barney, 2013, 2016; Barney & Fisher, 2016; Dawson, 2002, 2004; He, 2022; He & Kingsbury, 2016; Kingsbury, 2009; Pendrill 2019, 2024; Melin et al., 2021; Stenner et al., 2013; Stenner & Fisher, 2013; Williamson, 2018) of results demonstrating repeatable reproducibility of empirically stable and theoretically explained unit definitions set the stage for imagining, designing, and developing the kind of unit system Duncan has in mind. A major goal for us in compiling this book is simply to put this idea on the table as a serious matter for consideration.

The present volume, then, joins Duncan and many others (Cohen, 1994; Guttman, 1977, 1985; Meehl, 1967; Michell, 1986; Rodgers, 2010; Rogosa, 1987; Wilson, 2013a) in criticizing and offering alternatives to:

> the syndrome that I [Duncan] have come to call *statisticism*: the notion that computing is synonymous with doing research, the naive faith that statistics is a complete or sufficient basis for scientific methodology, the superstition that statistical formulas exist for evaluating such things as the relative merits of different substantive theories or the 'importance' of the causes of a 'dependent variable'; and the delusion that decomposing the covariations of some arbitrary and haphazardly assembled collection of variables can somehow justify not only a 'causal model' but also, praise the mark, a 'measurement model.' There would be no point in deploring such caricatures of the scientific enterprise if there were a clearly identifiable sector of social science research wherein such fallacies were clearly recognized and emphatically out of bounds. But in my discipline it just is not so. Individual articles of exemplary quality are published cheek-by-jowl with transpar-

> ent exercises in statistical numerology. If the muck were ankle deep, we could wade through it. When it is at hip level, our most adroit and most fastidious workers can hardly avoid getting some of it on their product. (Duncan, 1984b, pp. 226–227)

Focused new developments in the direction away from “statisticism” and toward consideration of the possible viability of extending the SI began in 2008. The record of events recounted above prompted initiatives aimed at possible collaborations with metrologists interested in transdisciplinary conceptual and operational overlaps with psychology and the social sciences.

1.4.3 New alliances

Upon investigation, it turned out that some metrologists (Beges et al., 2010; Berglund et al., 2012; Finkelstein, 1975, 1994, 2003, 2005; Mari, 2000, 2003, 2009; Mari et al., 2009; Mari & Sartori, 2007; Pendrill, 2008; Pendrill et al., 2010) had been seeking thought partners in psychology and the social sciences for years but had not identified additive conjoint models as a focal interest.

The International Measurement Confederation (IMEKO), whose membership is composed of globally distributed national metrology institutes, became the forum in which new alliances were formed. Conversations on measurement in physics and psychology ensued at Joint Symposia organized by the IMEKO Technical Committees on Measurement Science (TC7), Education and Training in Measurement and Instrumentation (TC1), and Measurements in Biology and Medicine (TC13), with the later addition of Measurements of Human Functions (TC18) (Fisher, 2008, 2010, 2012, 2014). At the 2008 IMEKO TC1-TC7-TC13 Joint Symposium held in Annecy, France, Ludwik Finkelstein opened a session (Finkelstein, 2008, 2009), introducing the next presentation (Fisher, 2009) and taking the opportunity to remark on a wide range of issues in the history and philosophy of science and measurement. Finkelstein noted that, in comparison with metrologists in the natural sciences, psychological measurement researchers had focused on model-based approaches to measurement on a broad scale for a longer time. He also expressed his personal opinion that, because natural scientists had the advantage of more thoroughly worked out theories, tools, methods, and standards, psychometricians in general were confronted with more difficult conceptual challenges than metrologists. He said that the formulation of measurement models setting forth clear inferential requirements for estimating interval quantities from ordinal observations was an important step forward in the advancement of measurement theory and practice, and that psychologists had made significant contributions deserving of increased attention in the natural sciences (Fisher, 2008).

At the 2010 IMEKO Joint Symposium hosted by Finkelstein, Sanowar Khan, Kenneth Grattan, and their colleagues in London, Finkelstein (2010, p. 2) said:

> The development of measurement science as a discipline has not paid adequate attention to the wider use of measurement. It is increasingly recognized that the wide range and diverse applications of measurement are based on common logical and philosophical principles and share common problems. However the concepts, vocabularies and methodologies in the various fields of measurement in the literature tend to differ. The development of a unified science of measurement appropriate for all domains of application seems to be desirable.

At that 2010 Joint Symposium, Luca Mari raised the question as to what it could mean for psychological and social instruments to be calibrated when calibration is always to a standard SI unit, which psychology and the social sciences do not have. That question immediately illuminated the nature of a mutually informative dialogue between the natural and social sciences (Fisher, 2010, p. 1279). On the one hand, the natural sciences lack, while psychology and the social sciences possess, decades of widely adopted traditions and norms concerning how ordinal observations can serve as a necessary and sufficient basis for estimating interval unit quantities and uncertainties. On the other hand, psychology and the social sciences lack the natural sciences' methods and expectations as to the value for communications, innovation, and commerce that stand to be obtained from distributed systems of instruments calibrated and metrologically traceable to quality assured unit standards.

Mari then suggested that a leader in psychological measurement modeling should be invited to give a special talk at the 2011 IMEKO Joint Symposium to be held in Jena, Germany (Scharff & Linß, 2011; Fisher, 2012a). Mark Wilson (2011, 2013b) gave that presentation and was backed up by several colleagues also presenting social and psychological measurement research applying Rasch's additive conjoint models (Bezruczko, 2011; Cano et al., 2011; Cooper & Fisher, 2011; Fisher, 2011; Fisher & Stenner, 2011; Granger & Bezruczko, 2011; Salzberger, 2011; Stenner et al., 2023). Several of these presentations were given in a session on fundamentals of measurement science chaired by Klaus-Dieter Sommer, one of the editors of the De Gruyter Series on Measurement Science in which this present volume appears.

Similar arrays of presentations on psychological and social measurement were given at subsequent IMEKO Joint Symposia in Genoa, Italy, in 2013, in Madeira in 2014, at the IMEKO World Congress in Prague in 2015, and at the 2016 Joint Symposium held at the University of California, Berkeley. This latter meeting was hosted by Wilson and Fisher (2016, 2018) and was the first such meeting attended by approximately equal numbers of social and natural scientists. At the following IMEKO World Congress in Belfast in 2018, a special session on psychological measurement was organized by Wilson and Fisher (2019), and significant participation by psychologists and social scientists continued at the 2017 Joint Symposium in Rio de Janeiro (Costa-Monteiro et al., 2018), Brazil; in St. Petersburg, Russia, in 2019 (Sapozhnikov & Taymanov, 2019); and in 2022 in Porto, Portugal (Benoit, 2022, 2023). Over the course of these recent years, Mari has productively collaborated with Wilson (Mari & Wilson, 2013, 2014) and Wilson's former students David Torres Irribarra and Andrew Maul (Mari et al., 2013, 2016, 2023; Maul et al., 2018, 2019; Wilson et al., 2015).

The first decade of the new millennium saw the emergence of MINET, a network audaciously entitled *Measuring the Impossible*, sponsored by the European Commission (Berglund et al., 2012; Pendrill, 2014; Pendrill et al., 2010). MINET brought together a multidisciplinary consortium of metrologists, physicists, engineers, psychophysicists, psychologists, and sociologists to address the challenges of measurements with persons from a metrological perspective. Keynote MINET researchers ranged from the BIPM Director Andrew Wallard to academics such as Giovanni Rossi (Genoa) and Birgitta Berglund (Karolinska) to well-known figures such as Ludwik Finkelstein, Fred Roberts, Damir Dzhafazov, and James Townsend. Parts of the MINET community subsequently merged with others, including IMEKO and IOMW (International Objective Measurement Workshop; Wright, 1992a, Wilson, 1992, 1994b; Engelhard & Wilson, 1996; Wilson et al., 1997; Wilson & Engelhard, 2000; Garner et al., 2010; Brown et al., 2011; Duckor et al., 2015). These associations have formed the basis of several collaborations (Cano et al., 2016, 2018a/b, 2019; Locoro et al., 2021; Melin et al., 2021; Pendrill & Fisher, 2013, 2015; Fisher et al., 2019), which have continued to bear fruit, as is evident in the production of several books in the *Springer Series in Measurement Science and Technology* (Fisher & Cano, 2023; Mari et al., 2023; Pendrill, 2019; Wilson & Fisher, 2017) as well as this volume.

1.5 Closing comments

Jeckelmann and Edelmaier (2023, p. 3) observe that "An extension of the concept of measurement will be necessary in future developments if the SI is to truly live up to its claim to be the universal language for all sciences." That extension need not, however, involve any further expansion of the oxymoronic concept of "ordinal quantity" or distinctions between "kinds of quantities" introduced in recent editions of the International Vocabulary of Measurement (JCGM, 2012, p. 15; Mari, 2009; Pendrill 2019). On the contrary, as was suggested by Finkelstein in his 2008 IMEKO talk in Annecy, it may be that the ongoing acceptance of ordinal scales for physical constructs such as hardness will soon give way to new representations in interval and ratio units. But an extended SI metrology for psychology and the social sciences will not be a mere elevation of those fields to a status akin to that of physics and the natural sciences. No, it would rather seem that exciting implications for a new art and science of complexity spanning the full range of fields are in store.

References

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science*. Princeton University Press.

Adams, R. J., Wu, M. L., Cloney, D., Berezner, A., & Wilson, M. (2020). ACER ConQuest: Generalised Item Response Modelling Software (Version 5.29). Australian Council for Educational Research. https://www.acer.org/au/conquest

Alder, K. (2002). *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. The Free Press.

Allen, D. D., & Pak, S. (2023). Improving clinical practice with person-centered outcome measurement. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person centered outcome metrology* (pp. 53–105). Springer.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, *65*(331), 1248–1255.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*(1), 69–81.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. North-Holland.

Andersen, E. B. (1999). Sufficient statistics in educational measurement. In G. N. Masters & J. P. Keeves (Eds.). *Advances in measurement in educational research and assessment* (pp. 122–125). Pergamon.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Andrich, D. (1988). *Sage University Paper Series on Quantitative Applications in the Social Sciences. Vol. series no. 07–068: Rasch models for measurement*. Sage Publications.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.), *Mathematical and Theoretical Systems: Proceedings of the 24th International Congress of Psychology of the International Union of Psychological Science, Vol. 4* (pp. 7–16). Elsevier Science Publishers.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, *75*(2), 292–308.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social, and health sciences*. Springer.

Andrich, D., Sheridan, B., & Luo, G. (2017). *RUMM 2030: Rasch unidimensional models for measurement*. RUMM Laboratory Pty Ltd [www.rummlab.com.au].

Aryadoust, S. V., Tan, H. A. H., & Ng, L. Y. (2019). A scientometric review of Rasch measurement: The rise and progress of a specialty. *Frontiers in Psychology*, *10*, 2197.

Ashworth, W. J. (2004). Metrology and the state: Science, revenue, and commerce. *Science*, *306*(5700), 1314–1317.

Barney, M. (2013). *Leading value creation: Organizational science, bioinspiration, and the cue see model*. Palgrave Macmillan.

Barney, M. (2016). Calibrating charisma: The many-facet Rasch model for leader measurement and automated coaching. *Journal of Physics: Conference Series*, *772*(012051), doi: 10.1088/1742-6596/772/1/012051

Barney, M., & Barney, F. (2024). Transdisciplinary measurement through AI: Hybrid metrology and psychometrics powered by Large Language Models. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. De Gruyter.

Barney, M., & Fisher, W. P., Jr. (2016). Adaptive measurement and assessment. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 469–490.

Beges, G., Drnovsek, J., & Pendrill, L. R. (2010). Optimising calibration and measurement capabilities in terms of economics in conformity assessment. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement*, *15*(3), 147–154.

Benoit, E. (2022). Editorial to selected papers from IMEKO TC1-TC7-TC13-TC18 Joint Symposium and MATHMET (European Metrology Network for Mathematics and Statistics) Workshops 2022. *Acta IMEKO*, *11*(4), 1–2.
Benoit, E. (2023). Editorial to selected papers from IMEKO TC1-TC7-TC13-TC18 joint symposium and MATHMET workshop 2022, 2nd part. *Acta IMEKO*, *12*(2), 1–2.
Berglund, B., Rossi, G. B., Townsend, J. T., & Pendrill, L. R. (Eds.). (2012). *Measurement with persons: Theory, methods, and implementation areas*. Psychology Press, Taylor & Francis Group.
Bezruczko, N. (Ed.). (2005). *Rasch measurement in health sciences*. JAM Press.
Bezruczko, N. (2011, September 2). Foundational imperatives for measurement with mathematical models. K.-D. Sommer (Chair), *Session on Fundamentals of measurement science*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology. https://www.db-thueringen.de/receive/dbt_mods_00019432
Black, M. (1946). *Critical thinking*. Prentice Hall.
Black, M. (1962). *Models and metaphors*. Cornell University Press.
Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 1–52.
Blok, A., Farias, I., & Roberts, C. (Eds.). (2020). *The Routledge companion to Actor-Network Theory*. Routledge.
Blok, A., Nakazora, M., & Winthereik, B. R. (2016). Infrastructuring environments. *Science as Culture*, *25*(1), 1–22.
Bode, R. K., Heinemann, A. W., & Semik, P. (2000). Measurement properties of the Galveston Orientation and Amnesia Test (GOAT) and improvement patterns during inpatient rehabilitation. *Journal of Head Trauma Rehabilitation*, *15*(1), 637–655.
Bohr, N. (1963). *Essays 1958–1962 on atomic physics and human knowledge*. John Wiley & Sons.
Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer.
Bowker, G. C. (2016). How knowledge infrastructures learn. In P. Harvey, C. B. Jensen, & A. Morita (Eds.). *Infrastructures and social complexity: A companion* (pp. 391–403). Routledge.
Bowker, G., Timmermans, S., Clarke, A. E., & Balka, E. (Eds.) (2015). *Boundary objects and beyond: Working with Leigh Star*. MIT Press.
Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, *42*, 631–634.
Brown, N., Duckor, B., Draney, K., & Wilson, M. (Eds.). (2011). *Advances in Rasch measurement, Vol. Two*. JAM Press.
Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research*, *1*(1), 8–22.
Bud, R., & Cozzens, S. E. (Eds.). (1992). *SPIE Institutes: Vol. 9. Invisible connections: Instruments, institutions, and science*. (R. F. Potter, Ed.). SPIE Optical Engineering Press.
Bulut, O. (2021). *Package 'eirm': Explanatory Item Response Modeling for Dichotomous and Polytomous Items*. University of Alberta, Edmonton: CRAN, https://github.com/okanbulut/eirm
Calude, C. S. (2002). Incompleteness, complexity, randomness and beyond. *Minds and Machines*, *12*, 503–517.
Calude, C. S. (Ed.). (2007). *Randomness & complexity from Leibniz to Chaitin*. World Scientific.
Cano, S., Klassen, A. F., & Pusic, A. L. (2011, September 2). *From Breast-Q © to Q-Score ©: Using Rasch measurement to better capture breast surgery outcomes*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology. http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24429/ilm1-2011imeko-039.pdf

Cano, S., Melin, J., Fisher, W. P., Jr., Stenner, A. J., & Pendrill, L., EMPIR NeuroMet 15HLT04 Consortium. (2018a). Patient-centred cognition metrology. *Journal of Physics: Conference Series*, *1065*, 072033. https://iopscience.iop.org/article/10.1088/1742-6596/1065/7/072033/meta

Cano, S., Pendrill, L., Barbic, S., & Fisher, W. P., Jr. (2018b). Patient-centred outcome metrology for healthcare decision-making. *Journal of Physics: Conference Series*, *1044*, 012057. http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012057

Cano, S., Pendrill, L., Melin, J., & Fisher, W. P., Jr. (2019). Towards consensus measurement standards for patient-centered outcomes. *Measurement*, *141*, 62–69. https://doi.org/10.1016/j.measurement.2019.03.056

Cano, S., Vosk, T., Pendrill, L., & Stenner, A. J. (2016). On trial: The compatibility of measurement in the physical and social sciences. *Journal of Physics: Conference Series*, *772*, 012025. doi: 10.1088/1742-6596/772/1/012025

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos*, *4*(1), 3–15.

Chien, T. W., Chang, Y., Wen, K. S., & Uen, Y. H. (2018). Using graphical representations to enhance the quality-of-care for colorectal cancer patients. *European Journal of Cancer Care*, *27*(1), e12591.

Chien, T.-W., Linacre, J. M., & Wang, W.-C. (2011). Examining student ability using KIDMAP fit statistics of Rasch analysis in Excel. In H. Tan & M. Zhou (Eds.), *Communications in Computer and Information Science: Vol. 201. Advances in Information Technology and Education, CSE 2011 Qingdao, China Proceedings, Part I* (pp. 578–585). Springer Verlag.

Choi, E. (1998). Rasch invents “ounces.” *Popular Measurement*, *1*(1), 29. https://www.rasch.org/pm/pm1-29.pdf

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186–190.

Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, *49*, 997–1003. https://psycnet.apa.org/record/1995-12080-001

Commons, M. L., Gane-McCalla, R., Barker, C. D., & Li, E. Y. (2014). The model of hierarchical complexity as a measurement system. *Behavioral Development Bulletin*, *19*(3), 9–14.

Connolly, A. J., Nachtman, W., & Pritchett, E. M. (1971). *Keymath: Diagnostic Arithmetic Test*. American Guidance Service, https://images.pearsonclinical.com/images/pa/products/keymath3_da/km3-da-pub-summary.pdf

Cooper, G., & Fisher, W. P., Jr. (2011, September 2). *Continuous quantity and unit; their centrality to measurement*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology. http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24494/ilm1-2011imeko-019.pdf.

Costa-Monteiro, E., Costa-Felix, R. P. B., & Barbosa, C. R. H. (2018). 2017 Joint IMEKO TC1-TC7-TC13 Symposium: “Measurement Science Challenges in Natural and Social Sciences”. *Journal of Physics Conference Series*, *1044*, (011001).

Crosby, A. W. (1997). *The measure of reality: Quantification and Western society, 1250-1600*. Cambridge University Press.

Dawson, T. L. (2002). New tools, new insights: Kohlberg’s moral reasoning stages revisited. *International Journal of Behavioral Development*, *26*(2), 154–166.

Dawson, T. L. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development*, *11*(2), 71–85.

Dawson, T. L., Fischer, K. W., & Stein, Z. (2006). Reconsidering qualitative and quantitative research approaches: A cognitive developmental perspective. *New Ideas in Psychology*, *24*, 229–239.

Dawson-Tunik, T. L., Commons, M., Wilson, M., & Fischer, K. (2005). The shape of development. *The European Journal of Developmental Psychology*, *2*, 163–196.
De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. (Statistics for Social and Behavioral Sciences). Springer-Verlag.
de Courtenay, N. (2015). The double interpretation of the equations of physics and the quest for common meanings. In O. Schlaudt & L. Huber *Standardization in Measurement* (pp. 53–68), London: Pickering & Chatto Publishers.
Dewey, J. (1929). *The quest for certainty: A study of the relation of knowledge and action*. Perigee Books, G. P. Putnam's Sons.
Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 package. *Journal of Statistical Software*, *20*(2), http://www.jstatsoft.org/v20/i02
Duckor, B., & Holmberg, C. (2019). Exploring how to model formative assessment trajectories of posing-pausing-probing practices: Toward a teacher learning progressions framework for the study of novice teachers. *Journal of Educational Measurement*, *56*(4), 836–890.
Duckor, B., Santelices, M. V., & Brandt, S. (2015). Advances in Rasch modeling: New applications and directions for objective measurement science. *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana*, *52*(2), 1–5. 10.7764/PEL.52.2.2015
Duncan, O. D. (1984a). Measurement and structure: Strategies for the design and analysis of subjective survey data. In C. F. Turner & E. Martin (Eds.). *Surveying subjective phenomena, Vol. 1* (pp. 179–229). Russell Sage Foundation.
Duncan, O. D. (1984b). *Notes on social measurement: Historical and critical*. Russell Sage Foundation.
Duncan, O. D. (1984c). Rasch measurement: Further examples and discussion. In C. F. Turner & E. Martin (Eds.). *Surveying subjective phenomena, Vol. 2* (pp. 367–403). Russell Sage Foundation.
Duncan, O. D. (1992). What if? *Contemporary Sociology*, *21*(5), 667–668.
Duncan, O. D., & Stenbeck, M. (1987). Are Likert scales unidimensional? *Social Science Research*, *16*(3), 245–259.
Duncan, O. D., & Stenbeck, M. (1988). Panels and cohorts: Design and model in the study of voting turnout. In C. C. Clogg (Ed.). *Sociological Methodology 1988* (pp. 1–35). American Sociological Association.
Edelmann, N. (2022). Digitalisation and developing a participatory culture: Participation, co-production, co-destruction. In Y. Charalabidis, L. S. Flak, & G. V. Pereira (Eds.). *Scientific foundations of digital governance and transformation (Public Administration and Information Technology, Vol. 38)* (pp. 415–435). Springer, https://doi.org/10.1007/978-3-030-92945-9_16
Einstein, A. (1954). *Ideas and opinions*. Bonanza Books.
Embretson, S. E. (1996a). Item Response Theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212.
Embretson, S. E. (1996b). The new rules of measurement. *American Psychologist*, *8*(4), 341–349.
Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.
Engelhard, G., Jr. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge Academic.
Engelhard, G., Jr. & Wilson, M. (1996). *Objective measurement: Theory into practice, Vol. 3*. Ablex.
Faber, M. (1999a). Numbers and psychoanalysis: Reflections on the quest for certainty, Part I of II. *Psychoanalytic Review*, *86*(1), 63–107. https://www.proquest.com/scholarly-journals/numbers-psychoanalysis-reflections-on-quest/docview/195060395/se-2
Faber, M. (1999b). Numbers and psychoanalysis: Reflections on the quest for certainty, Part II of II. *Psychoanalytic Review*, *86*(2), 245–279. https://www.proquest.com/docview/195066081/citation/2E9355EFE0154496PQ/1?accountid=135705

Finkelstein, L. (1975). Representation by symbol systems as an extension of the concept of measurement. *Kybernetes*, *4*(4), 215–223.
Finkelstein, L. (1994). Measurement and instrumentation science: An analytical review. *Measurement*, *14*(1), 3–14.
Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement*, *34*(1), 39–40 (10).
Finkelstein, L. (2005). Problems of measurement in soft systems. *Measurement*, *38*(4), 267–274.
Finkelstein, L. (2008, September 5). *Problems of widely-defined measurement*. Presented at the International Measurement Confederation (IMEKO) TC1 & TC7 Joint Symposium on Man, Science, & Measurement, Annecy, France.
Finkelstein, L. (2009). Widely-defined measurement–An analysis of challenges. *Measurement: Concerning Foundational Concepts of Measurement Special Issue Section*, *42*(9), 1270–1277.
Finkelstein, L. (2010). Measurement and instrumentation science and technology—the educational challenges. *Journal of Physics Conference Series*, *238*, 012001. doi: 10.1088/1742-6596/238/1/012001
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.
Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*(1), 59–77.
Fischer, G. H., & Molenaar, I. (1995). *Rasch models: Foundations, recent developments, and applications*. Springer-Verlag.
Fischer, K. W., & Dawson, T. L. (2002). A new kind of developmental science: Using models to integrate theory and research: Comment. *Monographs of the Society for Research in Child Development*, *67*(1), 156–167.
Fisher, R. A. (1935). The logic of inductive inference [with discussion]. *Journal of the Royal Statistical Society*, *98*(1), 39–82. https://doi.org/10.2307/2342435
Fisher, W. P., Jr. (1988). Truth, method, and measurement: The hermeneutic of instrumentation and the Rasch model [Diss]. *Dissertation Abstracts International (*Dept. of Education, Division of the Social Sciences, University of Chicago*), 49*, 0778A. (376 pages, 23 figures, 31 tables)
Fisher, W. P., Jr. (1992). Objectivity in measurement: A philosophical history of Rasch's separability theorem. In M. Wilson (Ed.). *Objective measurement: Theory into practice. Vol. I* (pp. 29–58). Ablex Publishing Corporation.
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. 118–153). Ashgate Publishing Co.
Fisher, W. P., Jr. (2003b). Mathematics, measurement, metaphor, metaphysics: Parts I & II. *Theory & Psychology*, *13*(6), 753–828.
Fisher, W. P., Jr. (2004). Meaning and method in the social sciences. *Human Studies: A Journal for Philosophy and the Social Sciences*, *27*(4), 429–454.
Fisher, W. P., Jr. (2008). Notes on IMEKO symposium. *Rasch Measurement Transactions*, *22*(1), 1147. http://www.rasch.org/rmt/rmt221.pdf
Fisher, W. P., Jr (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement*, *42*(9), 1278–1287.
Fisher, W. P., Jr. (2010). Unifying the language of measurement. *Rasch Measurement Transactions*, *24*(2), 1278–1281. http://www.rasch.org/rmt/rmt242.pdf
Fisher, W. P., Jr. (2011, September 1). Measurement, metrology and the coordination of sociotechnical networks. K.-D. Sommer (Chair), *Session on Fundamentals of measurement science*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology, http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24491/ilm1-2011imeko-017.pdf, Jena, Germany.

Fisher, W. P., Jr. (2012). 2011 IMEKO conference proceedings available online. *Rasch Measurement Transactions*, *25*(4), 1349. http://www.rasch.org/rmt/rmt254.pdf

Fisher, W. P., Jr. (2014). IMEKO 2008-2013 video presentations available online. *Rasch Measurement Transactions*, *27*(4), 1440–1441. https://www.rasch.org/rmt/rmt274b.htm

Fisher, W. P., Jr. (2019). A nondualist social ethic: Fusing subject and object horizons in measurement. *TMQ–Techniques, Methodologies, and Quality (Special Issue, Health Metrology)*, *10*, 21–40. https://portal2.ipt.pt/media/manager.php?src=servico&cmd=file&target=m1_MTc2NDQ#page=21

Fisher, W. P., Jr. (2020). Contextualizing sustainable development metric standards: Imagining new entrepreneurial possibilities. *Sustainability*, *12*(9661), 1–22. https://doi.org/10.3390/su12229661

Fisher, W. P., Jr. (2021). Bateson and Wright on number and quantity: How to not separate thinking from its relational context. *Symmetry*, *13*(1415), https://doi.org/10.3390/sym13081415

Fisher, W. P., Jr. (2023). Measurement systems, brilliant results, and brilliant processes in healthcare: Untapped potentials of person-centered outcome metrology for cultivating trust. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 357–396). Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr., & Cano, S. (Eds.). (2023). *Person-centered outcome metrology: Principles and applications for high stakes decision making*. Springer Series in Measurement Science & Technology. Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr., Oon, E. P.-T., & Benson, S. (2021). Rethinking the role of educational assessment in classroom communities: How can design thinking address the problems of coherence and complexity? *Educational Design Research*, *5*(1), 1–33. doi: 10.15460/eder.5.1.1537

Fisher, W. P., Jr, Pendrill, L., Lips da Cruz, A., & Felin, A. (2019). Why metrology? Fair dealing and efficient markets for the United Nations' Sustainable Development Goals. *Journal of Physics: Conference Series*, *1379*(012023), doi: 10.1088/1742-6596/1379/1/012023

Fisher, W. P., Jr., & Stenner, A. J. (2011, September 2). A technology roadmap for intangible assets metrology. In P. Scharff & G. Linß (Eds.). *International Measurement Confederation (IMEKO) TC1-TC7-TC13 Joint Symposium*, Jena, Germany, http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24493/ilm1-2011imeko-018.pdf, (Rpt. in W. P. Fisher, Jr., and P. J. Massengill (Eds.). (2023). *Explanatory models, unit standards, and personalized learning in educational measurement* (pp. 179–198). Springer, https://link.springer.com/book/10.1007/978-981-19-3747-7)

Fisher, W. P., Jr., & Wright, B. D. (Eds.). (1994). Applications of probabilistic conjoint measurement. *International Journal of Educational Research*, Vol. 21(6), 557–664.

Floyd, J., & Kanamori, A. (2016). Gödel vis-à-vis Russell: Logic and set theory to philosophy. In G. Crocco & E.-M. Engelen (Eds.). *Godelian Studies on the Max-Phil Notebooks, Vol. I* (pp. 243–326). Presses Universitaires de Provence.

Gadamer, H.-G. (1981). *Studies in Contemporary German Social Thought. Vol. 2: Reason in the age of science*. T. McCarthy (Ed.) (F. G. Lawrence, Trans.), MIT Press.

Gadamer, H.-G. (1989). *Truth and method*. (J. Weinsheimer & D. G. Marshall, Trans.) (Rev. ed.). Crossroad.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. University of Chicago Press.

Galison, P. (1999). Trading zone: Coordinating action and belief. In M. Biagioli (Ed.). *The science studies reader* (pp. 137–160). Routledge.

Garfinkel, A. (1991). Reductionism. In R. Boyd, P. Gasper, & J. D. Trout (Eds.). *The philosophy of science* (pp. 443–459). MIT Press.

Garner, M., Engelhard, G., Jr., Fisher, W. P., Jr., & Wilson, M. (Eds.). (2010). *Advances in Rasch measurement, Volume One*. JAM Press.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Lawrence Erlbaum Associates.

Goodman, L. A. (2007). Otis Dudley Duncan, quantitative sociologist par excellence: Path analysis, loglinear methods, and Rasch models. *Research in Social Stratification and Mobility*, *25*(2), 129–139. doi: 10.1016/j.rssm.2007.05.005

Granger, C., & Bezruczko, N. (2011). Body, mind, and spirit are instrumental to functional health: A case study. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology. https://www.db-thueringen.de/receive/dbt_mods_00019497

Green, K. E. (1986). Fundamental measurement: A review and application of additive conjoint measurement in educational testing. *Journal of Experimental Education*, *54*(3), 141–147.

Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, *12*(4), 369–381.

Guttman, L. (1977). What is not what in statistics. *The Statistician*, *26*, 81–107.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis*, *1*, 3–10.

Guttman, L. (1994). The mathematics in ordinary speech. In S. Levy (Ed.). *Louis Guttman on theory and methodology: Selected writings* (pp. 103–119). Dartmouth Publishing Company.

Hagell, P. (2014). Testing rating scale unidimensionality using the Principal Component Analysis (PCA)/t-test protocol with the Rasch model: The primacy of theory over statistics. *Open Journal of Statistics*, *4*(6), 456–465. doi: 10.4236/ojs.2014.46044

Hagell, P. (2019). Measuring Activities of Daily Living in Parkinson's disease: On a road to nowhere and back again? *Measurement*, *132*, 109–124. https://www.sciencedirect.com/science/article/pii/S0263224118308844

Haraway, D. J. (2022). A Cyborg Manifesto: An ironic dream of a common language for women in the integrated circuit. In *The Transgender Studies Reader Remix* (pp. 429–443). Routledge.

He, W. (2022). *MAP Growth item parameter drift study*. Portland, OR: NWEA Research Report.

He, W., Li, S., & Kingsbury, G. G. (2016). A large-scale, long-term study of scale drift: The micro view and the macro view. *Journal of Physics Conference Series*, *772*, 012022. https://iopscience.iop.org/article/10.1088/1742-6596/772/1/012022/meta

Hesse, M. (1970). *Models and analogies in science*. Notre Dame University Press.

Hohensinn, C. (2018). PcIRT: An R package for polytomous and continuous Rasch models. *Journal of Statistical Software*, *84*(2), 1–14. https://www.jstatsoft.org/article/view/v084c02/0

Holton, G. (1988). *Thematic origins of scientific thought: Kepler to Einstein* (revised ed.). Harvard University Press.

Huxley, T. H. (1862). *On our knowledge of the causes of the phenomena of organic nature*. Robert Hardwicke.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology*. (The Indiana Series in the Philosophy of Technology). Indiana University Press.

Jasanoff, S. (2004). *States of knowledge: The co-production of science and social order*. (International Library of Sociology). Routledge.

JCGM: Joint Committee for Guides in Metrology. Working Group 2, (2012). *International vocabulary of metrology: Basic and general concepts and associated terms, 3rd ed (with minor corrections)*, International Bureau of Weights and Measures–BIPM.

Jeckelmann, B., & Edelmaier, R. (Eds.). (2023). *Metrological infrastructure*. (K.-D. Sommer & T. Fröhlich, Eds.). (De Gruyter Series in Measurement Sciences). De Gruyter Oldenbourg, https://doi.org/10.1515/9783110715835

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475.

Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Karlsson, S., Svensson, H., Wisén, J., & Melin, J. (2024). A metrological approach to social sustainability metrics in municipalities. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 379–398), De Gruyter.

Kauffman, S. A., & Roli, A. (2023). A third transition in science? *Interface Focus*, *13*(3), 20220063.

Kielhofner, G., Dobria, L., Forsyth, K., & Basu, S. (2005). The construction of keyforms for obtaining instantaneous measures from the occupational performance history interview ratings scales. *OTJR: Occupation Participation and Health*, *25*(1), 23–32.

Kingsbury, G. G. (2009). Tools for measuring academic growth. *Journal of Applied Measurement*, *10*(1), 97.

Kline, M. (1980). *Mathematics: The loss of certainty*. Oxford University Press.

Knorr Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Kollias, C. (2024). Placing multiple panel cut scores on the same measurement scale. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 345–360), De Gruyter.

Kula, W. (1986). *Measures and men*. (R. Screter, Trans.), Princeton University Press, (Original work published 1970).

Kumar, D. N., & Joshi, H. B. (2024). Patient-centred outcome assessments in the surgical disciplines. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 449–472), De Gruyter.

Lamprianou, I. (2020). *Applying the Rasch model in social sciences using R and BlueSky statistics*. Quantitative Methodology Series. Routledge.

Lang, W. S., & Wilkerson, J. (2024). Measuring teacher dispositions: Steps in an innovative journey in affective assessment. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 301–344), De Gruyter.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Sage.

Lerner, R. M., & Overton, W. F. (2017). Reduction to absurdity: Why epigenetics invalidates all models involving genetic reduction. *Human Development*, *60*, 107–123. https://doi.org/10.1159/000477995

Li, Y. (2006). Using the open-source statistical language R to analyze the dichotomous Rasch model. *Behavioral Research Methods*, *38*(3), 532–541.

Linacre, J. M. (1997). Instantaneous measurement and diagnosis. *Physical Medicine and Rehabilitation State of the Art Reviews*, *11*(2), 315–324. http://www.rasch.org/memo60.htm

Linacre, J. M. (2023a). Advancing the metrological agenda in the social sciences. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 165–193). Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Linacre, J. M. (2023b). *A user's guide to FACETS Rasch-Model computer program, v. 3.85.1*. Winsteps.com, http://www.winsteps.com/a/facets-manual.pdf

Linacre, J. M. (2024). *A user's guide to WINSTEPS Rasch-Model computer program, v. 5.7.0*. Winsteps.com, https://www.winsteps.com/manuals.htm

Linacre, J. M., Chan, G., & Adams, R. J. (2013). An 'estimation bias' shootout in the wild West: CMLE, JMLE, MMLE, PMLE. *Rasch Measurement Transactions*, *27*(1), 1403–1405. http://www.rasch.org/rmt/rmt271.pdf

Liu, J. (2018). Development and translation of measurement findings for the motivation assessment for team readiness, integration, and collaboration self-scoring form. *American Journal of Occupational Therapy*, *72*(4_Supplement_1), 7211500015p1-7211500015p1.

Locoro, A., Fisher, W. P., Jr, & Mari, L. (2021). Visual information literacy: Definition, construct modeling and assessment. *IEEE Access*, *9*, 71053–71071. https://doi.org/10.1109/ACCESS.2021.3078429

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*(2), 143–155.

Luce, R. D. (1959). *Individual choice behavior*. Wiley.

Luce, R. D. (1978). Dimensionally invariant numerical laws correspond to meaningful qualitative relations. *Philosophy of Science*, *45*, 1–16.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, *31*, 19–26.

MacIntosh, R. (1998). Global attitude measurement: An assessment of the World Values Survey Postmaterialism Scale. *American Sociological Review*, *63*(3), 452–464.

Mallinson, T. (2024). Extending the justice-oriented, anti-racist framework for validity testing to the application of measurement theory in re(developing) rehabilitation assessments. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 399–426). De Gruyter.

Mari, L. (2000). Beyond the representational viewpoint: A new formalization of measurement. *Measurement*, *27*, 71–84.

Mari, L. (2003). Epistemology of measurement. *Measurement*, *34*(1), 17–30.

Mari, L. (2009). On (kinds of) quantities. *Metrologia*, *46*(3), L11–L15. https://doi.org/10.1088/0026-1394/46/3/N01

Mari, L., Lazzarotti, V., & Manzini, R. (2009). Measurement in soft systems: Epistemological framework and a case study. *Measurement*, *42*(2), 241–253.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2013). Quantification is neither necessary nor sufficient for measurement. *Journal of Physics Conference Series*, *459*(1), http://iopscience.iop.org/1742-6596/459/1/012007

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2016). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement*, *100*, 115–121. http://www.sciencedirect.com/science/article/pii/S0263224116307497

Mari, L., & Sartori, L. (2007). A relational theory of measurement: Traceability as a solution to the non-transitivity of measurement results. *Measurement*, *40*, 233–242.

Mari, L., & Wilson, M. (2013). A gentle introduction to Rasch measurement models for metrologists [abstract]. *Journal of Physics Conference Series*, *459*(012002), http://iopscience.iop.org/1742-6596/459/1/012002/pdf/1742-6596_459_1_012002.pdf

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327. http://www.sciencedirect.com/science/article/pii/S0263224114000645

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement, 2nd ed*. (Springer Series in Measurement Science and Technology). Springer, https://link.springer.com/book/10.1007/978-3-031-22448-5

Marmor, Y. N., & Bashkansky, E. (2020). Processing new types of quality data. *Quality and Reliability Engineering International*, *36*, 2621–2638. https://doi.org/10.1002/qre.2642

Massof, R. W. (2008). Editorial: Moving toward scientific measurements of quality of life. *Ophthalmic Epidemiology*, *15*, 209–211.

Massof, R. W., Bradley, C., & McCarthy, A. M. (2024). Constructing a continuous latent disease state variable from clinical signs and symptoms. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 269–300), De Gruyter.

Masters, G. N., & Keeves, J. P. (Eds.). (1999). *Advances in measurement in educational research and assessment*. Pergamon.

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018). The quality of measurement results in terms of the structural features of the measurement process. *Measurement*, *116*, 611–620.

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, *131*, 764–770. https://www.sciencedirect.com/science/article/pii/S026322411830808X?via%3Dihub

Maxwell, J. C. (1873). *A treatise on electricity and magnetism, Vol. 1*. Clarendon press.

Mead, R. J. (2009). The ISR: Intelligent Student Reports. *Journal of Applied Measurement*, *10*(2), 208–224.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Melin, J. (2024). Is validity a straightforward concept to be used in measurements in the human and social sciences? In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 157–190). De Gruyter.

Melin, J., Cano, S. J., Flöel, A., Göschel, L., & Pendrill, L. (2021). Construct specification equations: 'recipes' for certified reference materials in cognitive measurement. *Measurement: Sensors*, *18*, 100290. https://doi.org/10.1016/j.measen.2021.100290

Melin, J., & Pendrill, L. R. (2023). The role of construct specification equations and entropy in the measurement of memory. In W. P. Fisher Jr & S. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 269–309). Springer Series in Measurement Science and Technology. Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*, 398–407.

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8), 1024–1048.

Morell, L., Tan, S., & Wilson, M. (2024). Relationship between measures of "21st century skills" and the content underlying them. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 427–448). De Gruyter.

Moulton, M. (1993). Probabilistic mapping. *Rasch Measurement Transactions*, *7*(1), 268. http://www.rasch.org/rmt/rmt71b.htm

Nagel, E., & Newman, J. R. (1958). *Gödel's Proof*. New York University Press.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166–180.

Nersessian, N. J. (1996). Child's play. *Philosophy of Science*, *63*, 542–546.

Nersessian, N. J. (2008). *Creating scientific concepts*. MIT Press.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

Nielsen, C. R., & Lynch, G. (2022). *Gadamer's Truth and Method: A polyphonic commentary*. Rowman & Littlefield.

Overton, W. F. (1998). Developmental psychology: Philosophy, concepts, and methodology. *Handbook of Child Psychology*, *1*, 107–188.

Overton, W. F. (2002). Understanding, explanation, and reductionism: Finding a cure for Cartesian anxiety. In L. Smith & T. Brown (Eds.). *Reductionism and the development of knowledge* (pp. 29–51). Erlbaum.

Pelton, T., & Bunderson, V. (2003). The recovery of the density scale using a stochastic quasi-realization of additive conjoint measurement. *Journal of Applied Measurement*, *4*(3), 269–281.

Pendrill, L. R. (2008). Operating 'cost' characteristics in sampling by variable and attribute. *Accreditation and Quality Assurance*, *13*(11), 619–631.

Pendrill, L. R. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science*, *9*(4), 22–33. http://www.tandfonline.com/doi/abs/10.1080/19315775.2014.11721702

Pendrill, L. R. (2019). *Quality assured measurement: Unification across social and physical sciences*. (Springer Series in Measurement Science and Technology). Springer, https://link.springer.com/book/10.1007/978-3-030-28695-8

Pendrill, L. R. (2024). Quantities and units: Order amongst complexity. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 35–100). De Gruyter.

Pendrill, L. R., Emardson, R., Berglund, B., Gröning, M., Höglund, A., Cancedda, A., Quinti, G., Crenna, F., Rossi, G. B., Drnovsek, J., Gersak, G., Van der Heijden, G., Kallinen, K., & Ravaja, N. (2010).

Measurement with persons: A European network. *NCSLi Measure: The Journal of Measurement Science*, *5*(2), 42–54. https://doi.org/10.1080/19315775.2010.11721515

Pendrill, L., & Fisher, W. P., Jr (2013). Quantifying human response: Linking metrological and psychometric characterisations of man as a measurement instrument. *Journal of Physics Conference Series*, *459*, http://iopscience.iop.org/1742-6596/459/1/012057

Pendrill, L., & Fisher, W. P., Jr (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, *71*, 46–55. doi:http://dx.doi.org/10.1016/j.measurement.2015.04.010

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255.

Pesudovs, K. (2006). Patient-centred measurement in ophthalmology – A paradigm shift. *BMC Ophthalmology*, *6*(25).

Pesudovs, K. (2010). Item banking: A generational change in patient-reported outcome measurement. *Optometry and Vision Science*, *87*(4), 285–293.

Power, M. (2004). Counting, control, and calculation: Reflections on measuring and management. *Human Relations*, *57*(6), 765–783. doi:10.1177/0018726704044955

Prigogine, I. (1986). Science, civilization and democracy: Values, systems, structures and affinities. *Futures*, *18*(4), 493–507.

Prigogine, I. (1997). *The end of certainty: Time, chaos, and the new laws of nature*. Free Press.

Prigogine, I., & Stengers, I. (1984). *Order out of chaos: Man's new dialogue with nature*. Bantam Books.

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). Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Volume IV: Contributions to biology and problems of medicine* (pp. 321–333 [http://www.rasch.org/memo1960.pdf]). University of California Press.

Rasch, G. (1966a). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.). *Readings in mathematical social science* (pp. 89–108). Science Research Associates, https://www.rasch.org/memo19662.pdf

Rasch, G. (1966b). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19*, 49–57.

Robinson, M., Johnson, A. M., Walton, D. M., & MacDermid, J. C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Medical Research Methodology*, *19*(1), 36.

Robitzsch, A., Kiefer, T., & Wu, M., (2020). *TAM: Test Analysis Modules. R package version 3.5–19*. Salzburg, Austria: Federal Institute for Education Research, Innovation and Development of the Austrian School System, https://CRAN.R-project.org/package=TAM

Roche, J. (1998). *The mathematics of measurement: A critical history*. The Athlone Press.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1–12. https://doi.org/10.1037/a0018326

Rogosa, D. (1987). Casual [sic] models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics*, *12*(2), 185–195.

Salzberger, T. (2009). *Measurement in marketing research: An alternative framework*. Edward Elgar.

Salzberger, T. (2011). The quantification of latent variables in the social sciences: Requirements for scientific measurement and shortcomings of current procedures. K.-D. Sommer (Chair), *Session on Fundamentals of measurement science*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education* *and training. 31 August – 2 September*, *Jena, Germany.* Ilmenau University of Technology. https://www.db-thueringen.de/receive/dbt_mods_00019432

Sampson, G., Burgess, Y., McBride, N., & Steveno, B. (2024). A many-faceted measurement modeling approach for informing test specifications: Practical guidance from the national registry of emergency medical technicians. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 361–378). De Gruyter.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3 PL model. *Psychometrika*, *80*(2), 450–467.

San Martin, E., & Rolin, J. M. (2013). Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, *143*(1), 116–130.

San Martin, E., Perticará, M., Varas, I. M., Kenzo Asahi, K., & González, J. (2024). The role of identifiability in empirical research. In W. P. Fisher Jr. & L. R. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 131–156). De Gruyter.

Sapozhnikova, K., & Taymanov, R. (2019). Joint IMEKO TC1-TC7-TC13-TC18 Symposium, 2–5 July, St. Petersburg, Russia. *Journal of Physics: Conference Series*, *1379*, 011001. doi: 10.1088/1742-6596/1379/1/011001

Scharff, P., & Linß, G. (Eds.). (2011). *Intelligent Quality Measurements – Theory*, *Education and Training: Proceedings of the 14th Joint International IMEKO TC1+TC7+TC13 Symposium*, *31 Aug – 2 Sept., Jena, Germany.* Ilmenau University of Technology (13 pp.), https://www.db-thueringen.de/receive/dbt_mods_00018047?q=imeko

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Simon, H. A. (1997). *Models of bounded rationality: Empirically grounded economic reason Vol. 3*. MIT press.

Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society*, *1*, 25–39.

Smith, E. V., Jr, & Smith, R. M. (2004). *Introduction to Rasch measurement*. JAM Press.

Smith, E. V., Jr, & Smith, R. M. (2007). *Rasch measurement: Advanced and specialized applications*. JAM Press.

Smith, R. M. (1994). Reporting candidate performance on computer adaptive tests: IPARM. *Rasch Measurement Transactions*, *8*(1), 344–345.

Smith, R. M. (1996). Item component equating. In M. Wilson (Ed.). *Objective measurement: Theory into practice, Vol. 3* (pp. 289–308). Ablex Publishing Co.

Smith, R. M. (1997). The relationship between goals and functional status in the Patient Evaluation Conference System. *Physical Medicine and Rehabilitation: State of the Art Reviews*, *11*(2), 333–343.

Stenner, A. J., & Fisher, W. P., Jr (2013). Metrological traceability in the social sciences: A model from reading measurement. *Journal of Physics Conference Series*, *459*(012025), http://iopscience.iop.org/1742-6596/459/1/012025

Stenner, A. J., Fisher, W. P., Jr, Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement*, *4*(536), 1–14. doi:, 10.3389/fpsyg.2013.00536. (Rpt. in W. P. Fisher, Jr. & P. J. Massengill (Eds.). (2023). *Explanatory models, unit standards, and personalized learning in educational measurement* (pp. 223–250). Springer, https://link.springer.com/book/10.1007/978-981-19-3747-7?page=1#toc)

Stenner, A. J., Fisher, W. P., Jr., Stone, M. H., & Burdick, D. S. (2016). Causal Rasch models in language testing: An application rich primer. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 1–14). Springer.

Stenner, A. J., & Smith, M., III. (1982). Testing construct theories. *Perceptual and Motor Skills*, *55*, 415–426. (Rpt. in W. P. Fisher, Jr. & P. J. Massengill (Eds.). (2023). *Explanatory models, unit standards, and personalized learning in educational measurement* (pp. 31–42). Springer, https://link.springer.com/book/10.1007/978-981-19-3747-7?page=1#toc)

Stenner, A. J., Stone, M., & Burdick, D. (2023). How to model and test for the mechanisms that make measurement systems tick. In W. P. Fisher Jr. & P. J. Massengill (Eds.). *Explanatory models, unit* *standards, and personalized learning . . .* (pp. 199–211). Springer, https://link.springer.com/content/pdf/10.1007/978-981-19-3747-7_15?pdf=chapter%20toc, (Originally presented in session on Fundamentals of measurement science, K.-D. Sommer (Chair). International Measurement Confederation (IMEKO), Jena, Germany, 31 August 2011, https://www.db-thueringen.de/receive/dbt_mods_00019430)

Stephanou, A., & Fisher, W. P., Jr. (2013). From concrete to abstract in the measurement of length. *Journal of Physics Conference Series*, *459*, http://iopscience.iop.org/1742-6596/459/1/012026

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.). *Measurement: Definitions and theories* (pp. 18–63). Wiley.

Sul, D. (2024). Situating culturally specific assessment development within the disjuncture-response dialectic. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 473–498). De Gruyter.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J.-L., Slade, A., Lawton, G., Simone, A., Carter, J., Lundgren-Nilsson, A., Tripolski, M., Ring, H., Biering-Sorensen, F., Marincek, C., Burger, H., & Phillips, S. (2013). Assessing and adjusting for cross cultural validity of impairment and activity limitation scales through Differential Item Functioning within the framework of the Rasch model: The Pro-ESOR project. *Medical Care*, *42*(1), 137–148. http://www.jstor.org/stable/4640700

Teresi, J., Golden, R. R., Cross, P., Gurland, B., Kleinman, M., & Wilder, D. (1995). Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *Journal of Clinical Epidemiology*, *48*(4), 473–483.

Thurstone, L. L. (1959). Attitudes can be measured. In L. L. Thurstone, *The measurement of values* (pp. 215–233). University of Chicago Press, Midway Reprint Series. (Rpt. from Thurstone, L. L. (1928). *American Journal of Sociology*, *XXXIII*, 529–544.).

Torres Irribarra, D., & Freund, R. (2014). Wright map package for R. *Rasch Measurement Transactions*, *28*(1), 1456. http://www.rasch.org/rmt/rmt281.pdf

Toulmin, S. E. (1953). *The philosophy of science: An introduction*. Hutchinson's University Library.

Tukey, J. A. (1986). Data analysis and behavioural science. In L. V. Jones (Ed.), *The collected works of John A. Tukey, Volume III, Philosophy and principles of data analysis: 1949 – 1964* (pp. 187–390). Chapman & Hall.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.). *Rasch models: Foundations recent developments, and applications* (pp. 215–237). Springer.

Verhelst, N., Hatzinger, R., & Mair, P. (2007). The Rasch sampler. *Journal of Statistical Software*, *20*(4), http://www.jstatsoft.org/v20/i04

von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. Springer.

Weinsheimer, J. (1985). *Gadamer's hermeneutics: A reading of Truth and Method*. Yale University Press.

Weiss, D. J. (2021). *The Rasch model*. Retrieved 6 March 2024, from Assessment System Corp.: https://assess.com/the-rasch-model/.

Weitzel, T. (2004). *Economics of standards in information networks*. Physica-Verlag.

Williamson, G. L. (2018). Exploring reading and mathematics growth through psychometric innovations applied to longitudinal data. *Cogent Education*, *5*(1464424), 1–29.

Wilson, M. R. (1992). *Objective measurement: Theory into practice, Vol. 1*. Ablex.

Wilson, M. R. (1994a). Comparing attitude across different cultures: Two quantitative approaches to construct validity. In M. Wilson (Ed.). *Objective measurement: Theory into practice, Vol. 2* (pp. 271–294). Ablex.

Wilson, M. R. (1994b). *Objective measurement: Theory into practice, Vol. 2*. Ablex.

Wilson, M. R. (2011, September 1). The role of mathematical models in measurement: A perspective from psychometrics. L. Mari (Chair), *Plenary lecture*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September, Jena, Germany.* http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24178/ilm1-2011imeko-005.pdf, Jena, Germany.

Wilson, M. R. (2013a). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, *78*(2), 211–236.

Wilson, M. R. (2013b). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement*, *46*, 3766–3774. http://www.sciencedirect.com/science/article/pii/S0263224113001061

Wilson, M. R. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, *37*(1), 5–20.

Wilson, M. R. (2023). *Constructing measures: An item response modeling approach, 2nd ed*. Routledge, https://doi.org/10.4324/9781410611697

Wilson, M., & Engelhard, G. (2000). *Objective measurement: Theory into practice*, Vol. 5. Ablex Publishing.

Wilson, M., Engelhard, G., & Draney, K. (Eds.). (1997). *Objective measurement: Theory into practice, Vol. 4*. Ablex.

Wilson, M., & Fisher, W. P., Jr. (2016). Preface: 2016 IMEKO TC1-TC7-TC13 Joint Symposium: Metrology across the Sciences: Wishful Thinking? *Journal of Physics Conference Series*, *772*(1), 011001. http://iopscience.iop.org/article/10.1088/1742-6596/772/1/011001/pdf

Wilson, M., & Fisher, W. P., Jr. (Eds.). (2017). *Psychological and social measurement: The career and contributions of Benjamin D. Wright*. (Springer Series in Measurement Science and Technology.) Springer Nature, https://link.springer.com/book/10.1007/978-3-319-67304-2

Wilson, M., & Fisher, W. P., Jr. (2018). Preface of special issue Metrology across the Sciences: Wishful Thinking? *Measurement*, *127*, 577. https://doi.org/10.1016/j.measurement.2018.07.028

Wilson, M., & Fisher, W. P., Jr. (2019). Preface of special issue, psychometric metrology. *Measurement*, *145*, 190. https://www.sciencedirect.com/journal/measurement/special-issue/10C49L3R8GT

Wilson, M., Mari, L., Maul, A., & Torres Irribarra, D. (2015). A comparison of measurement concepts across physical science and social science domains: Instrument design, calibration, and measurement. *Journal of Physics Conference Series*, *588*(012034), http://iopscience.iop.org/1742-6596/588/1/012034

Wilson, M., Scalise, K., & Gochyyev, P. (2019). Domain modelling for advanced learning environments: The BEAR assessment system software. *Educational Psychology*, *39*(10), 1199–1217.

Wise, M. N. (1995). Precision: Agent of unity and product of agreement. Part III–"Today Precision Must Be Commonplace." In M. N. Wise (Ed.). *The values of precision* (pp. 352–361). Princeton University Press.

Wittgenstein, L. (1983). *Remarks on the foundations of mathematics*. G. H. von Wright, R. Rhees, & G. E. M. Anscombe (Eds.). (G. E. M. Anscombe, Trans.), MIT Press.

Wolfe, F., Van der Heijde, D. M., & Larsen, A. (2000). Assessing radiographic status of rheumatoid arthritis: Introduction of a short erosion scale. *Journal of Rheumatology*, *27*(9), 2090–2099.

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology*, *31*, 27–32.

Woolley, A. W., & Fuchs, E. (2011). Collective intelligence in the organization of science. *Organization Science*, *22*(5), 1359–1367.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems* (pp. 85–101). Educational Testing Service, http://www.rasch.org/memo1.htm

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. http://www.rasch.org/memo42.htm

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review*, *3*(1), 281–288. http://www.rasch.org/memo41.htm

Wright, B. D. (1988). Georg Rasch and measurement: Informal remarks by Ben Wright at the Inaugural Meeting of the AERA Rasch Measurement SIG, New Orleans – April 8, 1988. *Rasch Measurement Transactions*, *2*, 25–32. http://www.rasch.org/rmt/rmt23.htm

Wright, B. D. (1992a). The International Objective Measurement Workshops: Past and future. In M. Wilson (Ed.). *Objective measurement: Theory into practice, Vol. 1* (pp. 9–28). Ablex Publishing.

Wright, B. D. (1992b). Raw scores are not linear measures. *Rasch Measurement Transactions*, *6*(1), 208. http://www.rasch.org/rmt/rmt61n.htm

Wright, B. D. (1997a). Fundamental measurement for outcome evaluation. *Physical Medicine & Rehabilitation State of the Art Reviews*, *11*(2), 261–288.

Wright, B. D. (1997b). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52. http://www.rasch.org/memo62.htm, https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D. (2012). Benjamin D. Wright's annotated KeyMath Diagnostic Profile. *Rasch Measurement Transactions*, *25*(4), 1350. https://www.rasch.org/rmt/rmt254.pdf

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–867. http://www.rasch.org/memo44.htm

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press.

Wright, B. D., Mead, R. J., & Ludlow, L. H. (1980). *KIDMAP: person-by-item interaction mapping* (MESA Memorandum #29). Chicago: MESA Press [http://www.rasch.org/memo29.pdf]. (6 pp.)

Wright, B. D., & Stenner, A. J. (1998). Readability and reading ability. Paper Presented to the Australian Council on Education Research (ACER), Melbourne, Australia. (Rpt. in W. P. Fisher, Jr. & P. J. Massengill (Eds.). (2023). *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement* (pp. 89–107). Springer).

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. Wide Range, Inc, http://www.rasch.org/measess/me-all.pdf

Yumoto, F., & Stone, M. H. (2011). Comparing item calibration estimates using Winsteps and RUMM2010. *Rasch Measurement Transactions, 25*(3), 1337. https://www.rasch.org/rmt/rmt253e.htm