Springer Series in Measurement Science and Technology

Mark Wilson
William P. Fisher, Jr. *Editors*

Psychological and Social Measurement

The Career and Contributions of Benjamin D. Wright

Springer

Springer Series in Measurement Science and Technology

The Springer Series in Measurement Science and Technology comprehensively covers the science and technology of measurement, addressing all aspects of the subject from the fundamental principles through to the state-of-the-art in applied and industrial metrology, as well as in the social sciences. Volumes published in the series cover theoretical developments, experimental techniques and measurement best practice, devices and technology, data analysis, uncertainty, and standards, with application to physics, chemistry, materials science, engineering and the life and social sciences.

More information about this series at http://www.springer.com/series/13337

Mark Wilson • William P. Fisher, Jr.
Editors

Psychological and Social Measurement

The Career and Contributions of Benjamin D. Wright

Springer

Editors
Mark Wilson
Graduate School of Education
University of California, Berkeley
Berkeley, CA, USA

William P. Fisher, Jr.
Graduate School of Education
University of California, Berkeley
Berkeley, CA, USA

ISSN 2198-7807 ISSN 2198-7815 (electronic)
Springer Series in Measurement Science and Technology
ISBN 978-3-319-67303-5 ISBN 978-3-319-67304-2 (eBook)
https://doi.org/10.1007/978-3-319-67304-2

Library of Congress Control Number: 2017961728

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This volume is dedicated to Benjamin Drake Wright in recognition of his insight and scholarship, his dedication to measurement, and his unfailing support for his students and colleagues.

Acknowledgments

We would like to thank the Organizing Committee of the Festschrift in honor of Ben Wright, held at the Rehabilitation Institute of Chicago the weekend of Friday through Sunday, April 25–27, 2003: William P. Fisher, Jr. (chair), David Andrich, Kendon Conrad, George Engelhard, Jr., Allen Heinemann, Mary Lunz, Geoff Masters, Alan Tennant, Ev Smith, and Mark Wilson. Thanks to all of the contributors and participants at the conference.

We would like to thank those who presented at the book launch for this volume at an invitational symposium at the National Council for Measurement in Education annual conference in in San Antonio, Texas, in April 2017: by Karen Draney, George Engelhard, Jr., William P. Fisher, Jr., Mary Lunz, John Stahl, Mark Wilson, and Stefanie Wind.

We would also like to thank Denise Penrose, Senior Editor at Springer, for her enthusiasm, guidance, and patience.

All the photographs in this book were taken by the friends and family of Benjamin Wright and were provided by them to the editors for publication in this volume. We are grateful for their contributions.

Contents

The original version of this book was revised. An erratum to this book can be found at https://doi.org/10.1007/978-3-319-67304-2_16

Contributors

Ray Adams is Special Advisor, Global Educational Monitoring at the ACER, and Professorial Fellow of the University of Melbourne.

Trevor Bond is Adjunct Professor in the College of Arts, Society and Education at James Cook University, as well as in the Faculty of Education at Universiti Kebangsaan, Malaysia.

Ed Bouchard is a senior Alexander Technique (AT) teacher based in Chicago with over 30 years' experience in private lessons and academic classes.

Ryan P. Bowles is in Family Development & Human Studies at Michigan State University.

George Engelhard is Professor in the Department of Educational Psychology at the University of Georgia.

William P. Fisher, Jr. is a Research Associate with the BEAR Center at UC Berkeley, and consults independently via LivingCapitalMetrics.com.

Kevin J. Grimm is with the University of Arizona Department of Psychology.

Tracy L. Kline is a researcher at RTI International, Research Triangle Park, North Carolina.

Mike Linacre is the Research Director of Winsteps.com. From 1989 until 2001 he worked closely with Ben Wright in the MESA Psychometric Laboratory at the University of Chicago. Mike was a recipient of the Samuel J. Messick Memorial Lecture Award.

Mary Lunz was Director of Examination Activities for the Board of Registration of the American Society for Clinical Pathologists for many years, and then served as Executive Director of Measurement Research Associates, Inc. in Chicago.

Geoff Masters is Chief Executive Officer and a member of the Board of the ACER in Camberwell, Victoria, Australia. He was a recipient of the Samuel J. Messick Memorial Lecture Award.

Karen M. Schmidt is in the Department of Psychology, University of Virginia.

John Stahl is Principal Psychometrician, working on the NCLEX program with Pearson VUE in Chicago.

Gregory Stone is Associate Professor at The University of Toledo, and is Founder and Managing partner of MetriKs Amérique, LLC, a psychometric, evaluation and progressive educational consulting firm.

Mark Stone (retired) was a Core Professor, Director of Research, and Distinguished Service Professor at the Adler School of Professional Psychology in Chicago.

Herbert J. Walberg is distinguished visiting fellow at the Hoover Institution, University Scholar at the University of Illinois at Chicago, and has served on the National Assessment Governing Board.

Mark Wilson is Professor of Education at the University of California, Berkeley, a Fellow of AERA, a past president of the Psychometric Society and NCME, and was a recipient of the Samuel J. Messick Memorial Lecture Award.

Abbreviations

ACER	Australian Council for Educational Research
AERA	American Educational Research Association
APM	Applied Psychological Measurement
ASCP	American Society for Clinical Pathology
AT	Alexander Technique
BTD	Best Test Design
EPM	Educational and Psychological Measurement
IMEKO	International Measurement Confederation
IOMW	International Objective Measurement Workshop
IRT	Item Response Theory
JMLE	Joint Maximum Likelihood Estimation
JAM	Journal of Applied Measurement
JEM	Journal of Educational Measurement
JES	Journal of Educational Statistics
JOM	Journal of Outcome Measurement
MESA	Measurement, Evaluation, and Statistical Analysis concentration in the University of Chicago Department of Education
MOMS	Midwest Objective Measurement Seminars
NBME	National Board of Medical Examiners
OM:TiP	Objective Measurement: Theory into Practice book series
PM	Popular Measurement, published in four volumes through 2000
PROX	Algebraic approximation estimation algorithm
Rasch SIG	The AERA Rasch Measurement Special Interest Group
RMT	Rasch Measurement Transactions
SD	Semantic Differential
SRI	Social Research, Inc.
TAT	Thematic Apperception Test
UCON	Unconditional estimation algorithm; now known as JMLE

Chapter 1
Introduction to Benjamin Wright and His Contributions to Measurement Science

William P. Fisher, Jr. and Mark Wilson

Abstract In this chapter we briefly describe the facts of Ben Wright's professional career as a physicist and psychologist. We also make some perspective-setting remarks on the strengths and range of his accomplishments, on the nature of his engaging and sometimes-challenging personality, as well as on his perspicacity and forward-looking view on the roles of measurement in the scientific world. In doing so, we ask some questions about his career and work that we hope (and expect) are illuminated by the succeeding chapters of the volume. Of particular interest are the ways in which Wright drew from his deep experiences in physics, mathematics, computers, and psychoanalysis to set the stage for new advances in qualitative theory and quantitative precision in measurement science, advances that are proving to span a wide range of fields not limited to psychology and the social sciences. We also give some details of the original Conference that was the generator of many of the chapters in the Volume.

1.1 Remembering Ben Wright

In a career spanning more than five decades, Benjamin Drake Wright made foundational contributions to the theory and practice of measurement. His influence extends far beyond education and psychology, where his work in measurement began, into health care and the social sciences at large. Recent developments in measurement theory and practice connect Wright's work with physics and the natural sciences in ways that point in new directions for the future. Our goal in editing this volume is, then, not only to recognize and celebrate Ben Wright's accomplishments in psychology, education, and the social sciences, but also to introduce his work to scientists in other fields interested in issues of measurement

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-3-319-67304-2_16

W.P. Fisher, Jr. • M. Wilson (✉)
Graduate School of Education, University of California, Berkeley, Berkeley, CA, USA
e-mail: wfisher@berkeley.edu; markw@berkeley.edu

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_1

and technology. In presenting the biography and contributions of this seminal figure, we hope to engage this larger audience in the expanding conversation across the sciences about measurement and the communication of meaningful, transparent information.

Ben drew from a rich formative experience in his own education, beginning at the Little Red School House in Greenwich Village, continuing into philosophical studies with Max Black and in physics at Cornell, then to Bell Labs and the University of Chicago, working as an assistant to future Nobelists Charles H. Townes and Robert Mulliken. In the 1950s, searching for life beyond physics, Ben became a certified psychoanalyst working in Bruno Bettelheim's laboratory at the Orthogenic School, and met Georg Rasch in 1960 via his colleague, friend, and neighbour, the statistician L. J. Savage.

Over the course of his career, Ben could be simultaneously wide open to new ideas and dismissive of anything that struck him as nonsense. Comments and anecdotes found in the pages that follow show the range of Ben's human capacities from patience and care at one extreme to abrupt brush-offs at the other. Though he succeeded in providing dozens of students and colleagues with the tools for successful careers in a new field, some have wished he could have made fewer enemies of colleagues who could have been helpful allies. Despite these failings, Ben succeeded in influencing theory and practice on a broad scale.

Sometimes the extent of Ben's influence goes further than one might expect, given his reputation for being irascible. For instance, one of Wright's students (jointly with Darrell Bock), Robert Mislevy, holds an ETS chair named for Fred Lord, with whom Ben famously engaged in a verbal tussle at an American Educational Research Association (AERA) meeting in the early 1970s. In the same vein, Ron Hambleton offered sincere words of praise for Ben upon his passing (Royal, 2015), even though the two of them exchanged quite a few pointed barbs in their 1992 AERA debate on IRT (Hambleton, Wright, Crocker, Masters, & van der Linden, 1992).

The impact of Ben's teaching—and the quality of students he attracted and influenced—is evident in the positions his intellectual inheritors hold and the honors they have attained. For example, four of Ben's students (Mislevy, Linacre, Masters, and Wilson), have given the prestigious Samuel J. Messick Memorial Lecture at ETS. Ben himself was awarded the Association of Test Publishers Career Achievement Award in 2002, followed by Ron Hambleton in 2003, and by Ben's student, Betty Bergstrom, in 2015. Qualitatively meaningful and quantitatively rigorous models and methods of managing constructs measured with ability tests or rating scales are forever indebted in essential ways to Ben's contributions.

Ben's influence on measurement in education, health care, and many other fields continues to resonate around the world. Speaking at the University of Copenhagen in 2010 during the celebration of the 50th anniversary of the publication of Rasch's book, *Probabilistic Models for Some Intelligence and Attainment Tests,* Svend Kreiner, a student of Rasch's in Denmark, remarked on the fact that "none of us would be here speaking about the work of an obscure Danish mathematician were it not for Ben Wright." Similarly, at the 2012 Pearson Global Research Conference held in Fremantle, Western Australia, Peter Hill, CEO of the Australian Curriculum,

Assessment and Reporting Authority (ACARA), recalled hearing Ben Wright speak in Australia in the early 1980s on measurement technologies that still have not yet been brought fully into the light of day.

Ben addressed not only the technical demands of rigorous theory, models, estimation methods, software, instrument design, and validity assessment, but was also intimately involved in the development of cognitive models and predictive construct theories via his collaborations with Mark Stone on the Knox Cube Test and with Jack Stenner on the Lexile Framework. Ben also intuitively grasped the essential importance of professional collaboration, contributing to the formation of the Rasch Measurement Special Interest Group in the AERA, the Institute for Objective Measurement, and the International Objective Measurement Workshops, among other organizations and meetings.

Ben Wright was born March 30, 1926, in New York City, and he died in Chicago on October 25, 2015. He graduated from Cornell University in 1945 with honours in physics and philosophy, and completed a Ph.D. in the University of Chicago's Committee on Human Development in 1957. His entire career from 1947 on, spanning physics, psychology, and measurement, was spent at the University of Chicago. In a 1972 letter, Rasch (1988) remarked that "since his first visit to Denmark in 1964 BW [Ben Wright] has practiced an almost unbelievable activity in this field, and results have certainly not been lacking." As shown by the testimonials in this volume, that extraordinary vitality continued until Tuesday, October 30, 2001, when Ben suffered a cerebral accident related to an injury incurred in his youth. Though he recovered physically quite soon, he did not fully recover his intellectual capacities and passed away in 2015.

1.2 From Physicist to Psychoanalyst to Measurement Scientist

Ben often remarked on wanting to achieve measurement of a quality a physicist would accept. He found Rasch's solid grounding of measurement principles in a mathematical parameter separation theorem superior to the lack of foundations common in most statistical methods. Ben's experience working in physics alongside Nobelists like Townes and Mulliken also informed his attitude toward data: it had to make sense in terms of a model or lawful pattern of relationships. Describing accidental data using uninterpretable and overcomplicated models could not, by definition, in Ben's thinking, lead in productive directions.

In due course, Ben came to the opinion that "Today there is no methodological reason why social science cannot become as stable, as reproducible, and hence as useful as physics" (Wright, 1997, p. 44). Others have concurred, such as Andrich (1988, p. 22), who wrote 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." But far from advocating a mere imitation of physics in psychology and the social sciences, Wright's intensive

background in psychoanalysis, philosophy, science, and computers prepared him to be creative in highly original ways. He felt as equally at home in mathematics and software programs as he did in Freud's appropriations of the mythologies of Oedipus and Narcissus.

Consideration of Wright's background leads one to wonder how his interest in philosophy might have prepared him to grasp what Rasch had to say in 1960. Those familiar with Max Black's (1962) classic book, *Models and Metaphors*, may wonder whether the passage in it (pp. 226–227) on Maxwell's method of analogy figured in the philosophy class Wright took with Black as an undergraduate at Cornell. If so, did Wright recognize Rasch's (1960, pp. 110–115) application of Maxwell's treatment of Newton's second law as an example of that method of analogy? Was Wright familiar with the way Maxwell blended substantive qualitative insight into natural phenomena with mathematical rigor and reasoning? Did Ben try to do much the same thing in his work with Rasch's models for measurement?

Did Rasch ever mention to Wright that he spent considerable time in the company of scholars in a direct line of intellectual inheritance from Maxwell, economists trained in physics who were known for their enthusiastic application of Maxwell's method of analogy (Boumans, 1993, 2005; Fisher, 2010a, 2010b; Fisher & Stenner, 2013)? Was Wright aware that his friend and neighbor, the statistician L. J. Savage, had become acquainted with Rasch in 1947 in Chicago at the Cowles Commission for Research in Economics? Did Wright enact his own variation on Maxwell's method in the manner described by Nersessian (2002) as an extension of everyday model-based reasoning? Was Wright supportive of Fisher's (1988) dissertation study of metaphor at least in part because he understood the relevance of Black's (1993, p. 30) statement that "every metaphor is the tip of a submerged model"?

1.3 Ongoing Developments

These questions are, of course, unlikely to ever be answered, but they open up new and potentially productive lines of inquiry. For all its essential importance, methodological rigor alone is not sufficient to the task of improving the quality of measurement in education and the social sciences. In contrast with common practice these fields, researchers in physics, for instance, are not expected to design, create, calibrate, and maintain their own instruments or to be intimately involved in establishing the unit standards to which those instruments are traceable. One of the most important directions in which Ben's work may lead is toward a similar division of labor in psychology and the social sciences. Ben would have been delighted to have been part of developments over the last ten years or so involving new collaborations of psychometricians and weights and measures standards experts (metrologists) (Fisher & Stenner, 2016; Mari & Wilson, 2014; Pendrill, 2014; Pendrill & Fisher, 2015; Wilson, 2013; Wilson, Mari, Maul, & Torres Irribarra, 2015). This work may

well lead toward much-needed wider and deeper appreciations for the roles of theory and instrumentation in psychology and the social sciences.

In addition, there are strong indications that Wright's work will fundamentally impact not just psychology and the social sciences, but may contribute to a new conceptualization of the arts and sciences across the full range of disciplines, from physics to the humanities. This possibility was hinted at by Ludwik Finkelstein, an early leader in measurement philosophy, upon his first introduction to the works of Rasch and Wright. In a presentation to the 2008 IMEKO Joint Symposium in Annecy, France, Finkelstein observed that psychology and the social sciences have made more progress than the natural sciences in thinking through the measurement of complex constructs (Fisher, 2008). In remarks he offered at the 2010 instance of that meeting in London, Finkelstein said, "It is increasingly recognised that the wide range and diverse applications of measurement are based on common logical and philosophical principles and share common problems" (Finkelstein, 2010, p. 2).

In additional comments on the limits of the state of the art in physics, Finkelstein (2009) pointed out that even a physical attribute as commonplace and historical as hardness is not associated with a theory relating and reconciling different measurement approaches, such as the Mohs and Vickers tests. Finkelstein may have anticipated the possibility that applications of Rasch's and Wright's ideas on measurement would lead in the direction of new insights in this area. Several years later, Stone and Stenner (2014) sketched out preliminary results of the kind Finkelstein indicated had not yet been produced in physics. Though Cheng and Cheng (2004) apply dimensional analysis to hardness measurement with methods and aims quite similar to those associated with Rasch model applications, they do not attempt the integration of scales accomplished by Stone and Stenner (2014). Plainly, much remains to be done in this area.

There are further suggestions that psychology and the social sciences are pushing in this direction of a transformed quality of results. Pelton and Bunderson (2003) reconstruct a density scale from Rasch principles, just as Stephanou and Fisher (2013) recount their independently conceived and produced recoveries of linear length measures from ordinal observations. In the same vein, methods described by Wright (2000) were used to develop and test an explicit model and instrumentation for measuring heart failure severity as defined by the condition's clinical chemistry (Fisher et al., 2002; Fisher & Burton, 2010). Similar work has been done with the clinical chemistry of other conditions (Cipriani, Fox, Khuder, & Boudreau, 2005), and in creating a model of functional binocular vision (Powers, Fisher, & Massof, 2016). It may be that these studies and the principles and methods of psychometric modelling they demonstrate will serve as examples for new developments in theory and instrumentation not yet conceived in the natural sciences.

Working from a more theoretical level, Andrich's (2017) recent IMEKO Joint Symposium keynote in Rio de Janeiro shows that Rasch's stochastic approach to error and uncertainty provides a more fundamental basis for unit definitions than the natural sciences' deterministic approach. Taking up yet another perspective, Fisher (2017) situates Rasch measurement in the context of complex adaptive systems encompassing a wide range of natural and social phenomena. If these explorations

turn out to be productive directions for future research, they will owe enormous debts to the foresight, energy, imagination, and determination of Ben Wright.

1.4 Documenting Wright's Career Contributions

1.4.1 Personal Memories, Anecdotes, and Reflections

Chapters 2–15, by Wright's students and colleagues, recount biographical details and personal experiences dating from Mark Stone's and Herb Walberg's initial encounters with Ben in the 1950s. The chapters are not organized in chronological order but instead (hopefully) present a series of "imaginative variations on an invariant," to adopt Ricoeur's (1991, p. 196) phrase on the way we all equate life with the stories we can tell about it. Chapter 15 collects together brief memories, anecdotes, and comments from Wright on Rasch and the practice of science, and from others on Wright.

1.4.2 Three Early Wright Articles

Appendix A is the text of Wright's Sabbath Lecture, given at the New Experimental College in Thy, Denmark, in September 1967 (Wright, 1968). Appendices B and C are reprints of two of Wright's articles dating from 1958 and 1960. These articles are included here due to their clear indications of Wright's critical perspective on education and educational measurement in the years just before he met Rasch. Appendix B (Wright, 1960) explores the topics of what are today known as peer learning and formative practices, broadly conceived; the extended historical background and arguments Wright presents are highly relevant to today's concerns in education. Chapter 14 by Fisher makes some preliminary suggestions how these and other early articles by Wright relate to the broader context of his measurement work.

1.4.3 Students and Dissertations

The list of dissertation committees Ben chaired (Appendix D) has 65 entries; he served on another 53, for 118 in total. Measurement dissertation topics range from models to estimation to fit to writing assessments, report formats and applications across psychometrics, education, medicine, nursing, and other areas, including two involving aesthetic judgment.

Not all of the graduate students were at the University of Chicago; several were at the University of Illinois at Chicago, and at least one at Northwestern University

in Evanston (on which Ben served as an informal, but very involved, advisor). Most (74) of the dissertations were written by men, but a significant number (44) were by women.

1.4.4 Wright's Vita

Appendix E presents the entirety of Wright's CV. It may be that Ben had not finished updating it when his longstanding health problems ended his working years. The unnumbered entries, and those out of sequence, are shown as Ben left them.

1.4.5 Wright's Key Publications

Appendix F is an annotated bibliography of Wright's major measurement books and articles. For those new to Wright's work, or those interested in expanding their awareness of his ideas, these pieces are the primary points of entry into Wright's contributions to the mathematics of models, estimation, and fit; measurement theory; and the history and philosophy of science.

1.5 The 2003 Conference

On the weekend of April 26 and 27, 2003, a conference in honour of Ben was held at the Rehabilitation Institute of Chicago. The organizing committee that published a call for papers in *Rasch Measurement Transactions* (Vol. 16, No. 3, p. 885) included William P. Fisher, Jr. (Chair), David Andrich, Kendon Conrad, George Engelhard, Allen Heinemann, Mary Lunz, Geoff Masters, Alan Tennant, Everett Smith, and Mark Wilson. Proposals for that conference were invited to:

> address some aspect of the theme: "Access, Provocation, and the Development of Professional Identity: Celebrating the Careers of Benjamin D. Wright." Though the choice of the specific topics addressed is for you to make, we hope that you will take up an issue that involves or builds on Ben's extensive contributions to making measurement more accessible and to the fundamental foundations of measurement, his reputation as an irascible provocateur, his selfless support for others' professional development, and/or his multiple careers, as explained below."

The sub-themes of the conference were elaborated thus:

> Access to Measurement: simpler, faster estimation (PROX, UCON); software that works; models for more kinds of data; error, reliability, and fit statistic development; publishing (MESA Press, RMT, support for OM:TiP, JOM, JAM, PM); associations (the SIG, IOM); meetings (MOMS, AERA/SIG, IOMW); and constant improvement to all of that via substantive interactions with students and colleagues.

Foundations of Measurement: measurement as a scientific enterprise, relation to scientific revolutions, relation to foundational ideas such as specific objectivity and additive conjoint measurement, relation to foundational work of figures such as Thurstone, Guttman and Rasch.

Provocation of and the Development of Professional Identity: Ben is well-known for strongly challenging and even abruptly dismissing anything that strikes him as irrelevant, foolish, or half-baked, and he seems to have had explicit reasons for behaving in this manner, reasons stemming from his work on identity development with Bruno Bettelheim. Personal accounts of Ben's successes and failures in this regard are of particular interest.

Multiple Careers: In addition to his work in measurement theory and practice, Ben worked as a physicist, and then as a psychologist and factor analyst. He taught a course on the psychology of becoming a teacher for many years, and continued working in this area long after most people associated him primarily with Rasch measurement. Even within the area of measurement alone, Ben's early work on estimation, models, fit, error, reliability, and software stands in considerable contrast with his later emphases on applications, organizations, and publishing. Papers touching on more than one of these careers will be of special interest.

A full list of the presenters and titles at the Conference appears in a later issue of *Rasch Measurement Transactions* (Vol. 17, No. 1, pp. 908–909).

In this volume, we have included a selection of papers from that conference that focused on Ben Wright's personal history, his character, and/or his accomplishments. We have added some invited chapters as well, covering some particular aspects of those. There was one paper presented at the 2003 conference we would have liked to include, but could not, as it was previously published (Andrich, 2015). A more voluminous collection of research reports, also based on the presentations from the original Conference, documenting Ben's influence on modeling, estimation, data quality evaluation, software, and applications, is in preparation.

References

Andrich, D. (1988). *Sage University Paper Series on quantitative applications in the social sciences. Vol. series no. 07–068: Rasch models for measurement*. Beverly Hills, California: Sage Publications.

Andrich, D. (2015). Ben Wright: "Idiosyncrasies of Autobiography and Personality" in taking up the Rasch measurement paradigm. *Rasch Measurement Transactions, 29*(3), 1539–1542.

Andrich, D. (2017). A law of ordinal random error: The Rasch measurement model and random error distributions of ordinal assessments. *Journal of Physics Conference Series*. in press.

Black, M. (1962). *Models and metaphors*. Ithaca, New York: Cornell University Press.

Black, M. (1993). More about metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 19–43). Cambridge: Cambridge University Press. (Reprinted from Black, M. (1977). More about metaphor. *Dialectica, 31*(3–4), 431–457).

Boumans, M. (1993). Paul Ehrenfest and Jan Tinbergen: A case of limited physics transfer. In N. De Marchi (Ed.), *Non-natural social science: Reflecting on the enterprise of "More Heat than Light"* (pp. 131–156). Durham, NC: Duke University Press.

Boumans, M. (2005). *How economists model the world into numbers*. New York: Routledge.

Cheng, Y. T., & Cheng, C. M. (2004). Scaling, dimensional analysis, and indentation measurements. *Materials Science and Engineering: R: Reports, 44*(4), 91–149.

Cipriani, D., Fox, C., Khuder, S., & Boudreau, N. (2005). Comparing Rasch analyses probability estimates to sensitivity, specificity and likelihood ratios when examining the utility of medical diagnostic tests. *Journal of Applied Measurement, 6*(2), 180–201.

Finkelstein, L. (2009). Widely-defined measurement: An analysis of challenges. *Measurement, 42*(9), 1270–1277.

Finkelstein, L. (2010). Measurement and instrumentation science and technology-the educational challenges. *Journal of Physics Conference Series, 238*, 012001.

Fisher, W. P., Jr. (1988). Truth, method, and measurement: the hermeneutic of instrumentation and the Rasch model, diss. Dissertation Abstracts International 49, 0778A. Dept. of Education, Division of the Social Sciences: University of Chicago. 376 pages, 23 figures, 31 tables.

Fisher, W. P., Jr. (2008). Notes on IMEKO symposium. *Rasch Measurement Transactions, 22*(1), 1147.

Fisher, W. P., Jr. (2010a). Rasch, Maxwell's method of analogy, and the Chicago tradition. Paper presented at the conference, Celebrating 50 years since the publication of Rasch's Probabilistic Models, University of Copenhagen School of Business, FUHU Conference Centre, Copenhagen, Denmark.

Fisher, W. P., Jr. (2010b). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics Conference Series, 238*(1), 012016.

Fisher, W. P., Jr. (2017). A practical approach to modeling complex adaptive flows in psychology and social science. *Procedia Computer Science*, *114*, 165–174.

Fisher, W. P., Jr., Bernstein, L. H., Qamar, A., Babb, J., Rypka, E. W., & Yasick, D. (2002). At the bedside: Measuring patient outcomes. *Advance for Administrators of the Laboratory, 11*(2), 8. 10.

Fisher, W. P., Jr., & Burton, E. (2010). Embedding measurement within existing computerized data systems: Scaling clinical laboratory and medical records heart failure data to predict ICU admission. *Journal of Applied Measurement, 11*(2), 271–287.

Fisher, W. P., Jr., & Stenner, A. J. (2013). On the potential for improved measurement in the human and social sciences. In Q. Zhang & H. Yang (Eds.), *Pacific Rim Objective Measurement Symposium 2012 Conference Proceedings*. Berlin, Germany: Springer-Verlag.

Fisher, W. P., Jr., & Stenner, A. J. (2016). Theory-based metrological traceability in education: A reading measurement network. *Measurement, 92*, 489–496.

Hambleton, R., Wright, B. D., Crocker, L., Masters, G., & van der Linden, W. (1992). Hambleton's 9 theses. *Rasch Measurement Transactions, 6*(2), 215.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement, 51*, 315–327.

Nersessian, N. J. (2002). Maxwell and "the method of physical analogy:" Model-based reasoning, generic abstraction, and conceptual change. In D. Malament (Ed.), *Reading natural philosophy: Essays in the history and philosophy of science and mathematics* (pp. 129–166). Lasalle, Illinois: Open Court.

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. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science, 9*(4), 22–33.

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement, 71*, 46–55.

Powers, M., Fisher, W. P., Jr. & Massof, R. W. (2016). Modeling visual symptoms and visual skills to measure functional binocular vision. *Journal of Physics Conference Series, 772*(1), 012045.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen and Lydiche. [Republished, 1980, University of Chicago Press.]

Rasch, G. (1988/1972). Review of the cooperation of Professor B. D. Wright, University of Chicago, and Professor G. Rasch, University of Copenhagen; letter of June 18, 1972. *Rasch Measurement Transactions, 2*(2), 19.

Ricoeur, P. (1991). Narrative identity. In D. Wood (Ed.), *On Paul Ricoeur: Narrative and interpretation* (pp. 188–199). New York, New York: Routledge.

Royal, K. (Ed.). (2015). A tribute to Benjamin D. Wright [Special issue]. *Rasch Measurement Transactions, 29*(3), 1528–1546.

Stephanou, A., & Fisher, W. P., Jr. (2013). From concrete to abstract in the measurement of length. *Journal of Physics Conference Series, 459*, 012026.

Stone, M. H., & Stenner, A. J. (2014). From ordinality to quantity. *Rasch Measurement Transactions, 27*(4), 1439–1440.

Wilson, M. R. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46*, 3766–3774.

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.

Wright, B. D. (1960). Should children teach? *The Elementary School Journal, 60*, 353–369.

Wright, B. D. (1968). The sabbath lecture: Love and order. In A. R. Nielsen et al. (Eds.), *Lust for learning*. Thy, Denmark: New Experimental College Press.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45, 52.

Wright, B. D. (2000). Rasch regression: My recipe. *Rasch Measurement Transactions, 14*(3), 758–759.

Chapter 2
Cogitations on Invariant Measurement

A Memo to Ben Wright on the Perspectives of Rasch and Guttman

George Engelhard, Jr.

Abstract The purpose of this chapter is to trace the evolution of the concept of invariant measurement as it appears in the work of Guttman and Rasch. The first section of the paper describes the concept of invariance. This section includes a detailed description of the perspectives of Guttman and Rasch on invariant measurement. The next section presents a re-analysis of the Stouffer-Toby data set using Guttman scaling and Rasch measurement theory. Finally, the implications for research, theory and practice of measurement are discussed. An earlier version of this research was presented at the Ben Wright Conference in Chicago (April 2003). I dedicate this chapter to Ben because it represents a continuation of my memos to him on the foundational ideas of measurement.

2.1 Personal Note

In the summer of 1977, I first met Professor Benjamin D. Wright at the University of Chicago when I applied to the MESA (Measurement, Evaluation, and Statistical Analysis) program. Ben shared with me a copy of his ETS paper (Wright, 1968), an article by Bruce Choppin (1968), and several publications by Georg Rasch (1961, 1966). A few years later, I enrolled in Ben's seminar on Psychometric Theory and this course required a reading log that he called *cogitations*. He described one of the assignments as follows:

G. Engelhard, Jr. (✉)
University of Georgia, Atlanta, GA, USA
e-mail: gengelh@uga.edu

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_2

Part A. 8 weekly memos containing:

1. An interesting quote on the requirements and/or methods of measurement. (These quotes should come about half from Thurstone, Guttman, Loevinger, Torgerson, etc., and about half from articles in current issues of APM, JEM, JES, EPM, and PM)
2. Your own *comments* on this quote.

Part B. A thoughtful *essay* drawn from your memos. (10–15 pages typed, signed, and dated).

This assignment captivated me, and started me on a journey to explore the history of ideas related to measurement theory and practice. In particular, my class readings and memos laid the ground work for my views of invariant measurement, as well as a philosophical perspective on many of the problems encountered in educational and psychological measurement (Engelhard, 2013).

2.2 The Concept of Invariant Measurement

Science is impossible without an evolving network of stable measures.
(Wright, 1997, p. 33)
The scientist is usually looking for invariance whether he knows it or not
(Stevens, 1951, p. 20)

The quest for stable measures has a long history in mathematics and science, and Wright (1997) has argued persuasively for its emergence in trade and construction. In fact, Wright viewed the stability of measures as a *moral necessity.* The Harvard philosopher Nozick (2001) has stressed that "evolution has shaped our sensory apparatus to give us the capacity to detect invariances [in our environment]" (p. 78). It is clear that invariance is a fundamental and guiding concept in numerous human activities.

In seeking stable measures, Ben Wright identified the requirements for objective measurement:

> First, the calibration of measuring instruments must be independent of those objects that happen to used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for the measuring (Wright, 1968, p. 87).

The first part of this quote refers to *person-invariant item calibration.* The basic measurement problem addressed by sample-invariant item calibration is how to minimize the influence of arbitrary samples of individuals on the estimation of item scale values or item difficulties. The overall goal of person-invariant measurement can be viewed as estimating the locations of items on a latent variable or construct of interest.

The second part of the quote refers to *item-invariant measurement of persons*. In the case of item-invariant measurement, the basic measurement problem is to minimize the influences of the particular items that happen to used to estimate a person's location on the latent variable or construct (Engelhard, 1984). Overall, both item and person locations should remain stable and consistent across various subsets of items and subgroups of persons. The final outcome of a successful measurement process is to create a unidimensional Wright Map to represent the latent variable.

Stevens (1951) presented a very strong case for the importance of the concept of invariance. He argued that "many psychological problems are already conceived as the deliberate search for invariances" (Stevens, 1951, p. 20). In developing his views of invariant measurement, Stevens was influenced by the insightful work of Mosier (1940, 1941) who pointed out the symmetry between psychophysics and psychometrics. According to Stevens (1951),

> psychophysics sees the response as an indicator of an attribute of the individual—an attribute that varies with the stimulus and is relatively invariant from person to person. Psychometrics regards the response as indicative of an attribute that varies from person to person but is relatively invariant for different stimuli. Both psychophysics and psychometrics make it their business to display the conditions and limits of invariances (p. 31).

Measurement problems related to invariance can viewed in terms of this duality between person-invariant item calibration and item-invariant person measurement. Invariant measurement reflects several key requirements that are necessary for successful measurement in the social, behavioral and health sciences. The quest for invariant measurement within the measurement research of Guttman (Engelhard, 2005) is described next.

2.3 Guttman Scaling

> *[Guttman scaling] displays in a rudimentary form virtually all the major properties and problems that characterize the more general scaling models.*
> (Mokken, 1971, p. 24).

Guttman (1944, 1950) laid the groundwork for a new technique designed to scale a set of items. According to Guttman (1950),

> One of the fundamental problems facing research workers … is to determine if the questions asked on a given issue have a single meaning for the respondents. Obviously, if a question means different things to different respondents, then there is no way that the respondents can be ranked … Questions may appear to express a single thought and yet not provide the same kind of stimulus to different people (p. 60).

In essence, Guttman is seeking a unidimensional representation of a construct with invariant and stable meaning across persons.

Guttman Scaling provides a framework for determining whether or not a set of items and group of persons met of the requirements of a Guttman Scale. Guttman determined the requirements of these perfect scales based on his view of an *ideal scale*.

According to Guttman, an ideal or perfect scale exists when person scores reproduce the exact item responses in the data matrix. In his words,

> A particularly simple representation of the data would be to assign to each individual a numerical value and to each category of each attribute a numerical value such that, given the value of the individual and the values of the categories of an attribute, we could reproduce the observations of the individual on the attribute (Guttman, 1944, p. 143).

One popular graphical method for identifying an ideal scale is called a scalogram. When persons and items are ordered and displayed in a table, then the data matrix forms a distinctive triangular pattern. For example, an ideal scale for four items (A, B, C, and D) has the following pattern:

	Four Items			
Person scores	A	B	C	D
4	1	1	1	1
3	1	1	1	0
2	1	1	0	0
1	1	0	0	0
0	0	0	0	0

These five item response patterns are ideal from Guttman's perspective because if we know the person's score, then we know exactly which items are answered correctly or incorrectly. There are 16 possible response patterns ($2^4 = 16$) for four items scored dichotomously, and only these five patterns define a Guttman scale.

In addition to seeking invariant meaning across persons in their interpretation of items, Guttman added the following requirement:

> If a person endorses a more extreme statement, he should endorse all less extreme statements if the statements are to be considered a scale … We shall call a set of items of common content a scale if a person with a higher rank than another person is just as high or higher on every item than the other person (Guttman, 1950, p. 62)

This also conforms to the concept of conjoint transitivity (Wright, 1997).

It is debatable whether or not Guttman used the idea of a latent variable; however, item response functions can be used to represent Guttman scaling. Figure 2.1 illustrates a perfect Guttman scale with four item response functions. The *x*-axis in Fig. 2.1 represents the latent variable with items A to D ordered from easy to hard. The items are rank ordered, and there is no requirement that the *x*-axis have equal intervals. The *y*-axis represents the probability of responding with a correct answer to each item. For example, a person located on the *x*-axis between items B and C is expected to answer Items A and B correctly (probability is 1.00) and items C and D incorrectly (probability is .00); this yields a score of 2 and a perfect response pattern [1100]. These distinctive step functions also serve to define a Guttman scale.

There are several methods proposed for examining model-data fit to a Guttman scale. The methods proposed for evaluating Guttman scales involve a comparison between observed and ideal response patterns. Guttman (1944) recognized that "perfect scales are not to be expected in practice" (p. 140), and that "deviation from

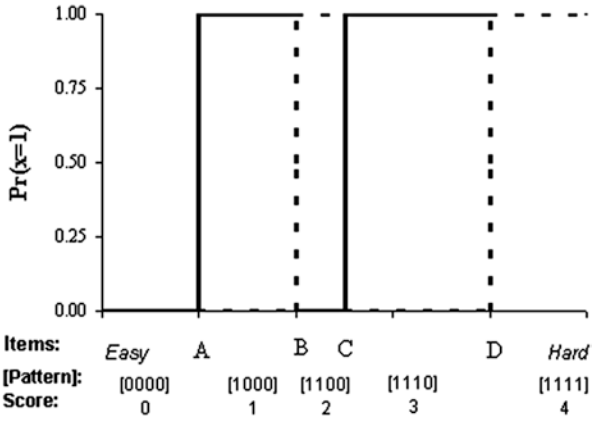


Fig. 2.1 Illustrative item response functions for Guttman Items (Deterministic model)

perfection is measured by a coefficient of reproducibility" (p. 140). Guttman originally proposed a coefficient of reproducibility as follows:

> The amount by which a scale deviates from the ideal scale patterns is measured by a *coefficient of reproducibility* [italics in the original] … It is secured by counting up the number of responses which would have been predicted wrongly for each person on the basis of his scale score, dividing these errors by the total number of responses and subtracting the results fraction from 1 … An acceptable approximation to a perfect scale has been arbitrarily set at 90 per cent reproducibility (Guttman, 1950, p. 77).

This coefficient of reproducibility is defined as:

$$\text{Rep} = 1 - \text{total number of errors} / \text{total number of responses} = 1 - [E / nk] \qquad (2.1)$$

where E represents an error count, n is the number of persons and k is the number of items. Engelhard (2008) provides a description of other model-data fit indices for Guttman scales that includes Loevinger's coefficient of homogeneity (Loevinger, 1947, 1948). Loevinger's coefficient of homogeneity is of particular importance because it was incorporated and extended by Mokken and others into nonparametric item response theory.

There are several aspects of Guttman scaling that should be noted. First of all, it is a deterministic model. Second, it is an ideal-type model that specifies clearly the desirable properties of a measure, and then examines model-data fit to see if the data conform to the measurement requirements. The quest for invariant measurement within the measurement research of Rasch is described next.

2.4 Rasch Measurement Theory

> *One of the best introductions to this change of paradigm is Rasch (1960, Chapter 1), which is mandatory reading …*
> (van der Linden, 2016, p. xiii)

Educational and psychological measurement in the first half of the twentieth century was dominated by what I have called the Test Score Tradition (Engelhard, 2013). As its label suggests, the Test Score Tradition is dominated by sum scores with Classical Test Theory as a key example of measurement research in this tradition (Crocker and Algina, 1986). The second half of the 20th century witnessed the emergence of a Scaling Tradition that recognized the duality between items and person scores (Mosier, 1940, 1941).

As pointed out by van der Linden (2016), Rasch was one of the pioneers within the tradition that represented a paradigm shift from earlier measurement research. Rasch (1960), presented a set of ideas and methods described by Loevinger (1965) as a "truly new approach to psychometric problems" (p. 151) that can lead to "non-arbitrary measures" (p. 151). Rasch sought to develop "individual-centered statistical techniques [that] require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated" (Rasch, 1960, p. xx).

Problems of invariant measurement played a central role in the development of Rasch's measurement theory. As pointed out by Andrich (1988), Rasch presented "two principles of invariance for making comparisons that in an important sense precede though inevitably lead to measurement" (p. 18). Problems related to invariance played a key role in motivating his measurement theory. Rasch's concept of specific objectivity and his principles of comparison form his version of the requirements for invariant measurement (Rasch, 1977). In his words,

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which stimuli within the considered class were or might also have been compared . Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should be independent of which other individuals were also compared, on the same or on some other occasion (Rasch, 1961, pp. 331–332).

It is clear in this quotation that Rasch recognized the importance of both person-invariant item calibration, and item-invariant measurement of persons. In fact, he made them cornerstones in his quest for specific objectivity. In order to address problems related to invariance, Rasch laid the foundation for the development of a family of measurement models that are characterized by the potential to separate item and person parameters (Wright & Masters, 1982).

Andrich (1985) has made a strong and persuasive case for viewing the Rasch model as a probabilistic realization of a Guttman scale. Rasch measurement theory can be used to model the probability of dichotomous item responses as a logistic function of item difficulty and person location on the latent variable. The dichotomous Rasch model can be written as follows:

$$\Pr(x_{ni} = 1|\theta_n, \delta_i) = \exp(\theta_n - \delta_i) / [1 + \exp(\theta_n - \delta_i)] \tag{2.2}$$

and

$$\Pr(x_{ni} = 0|\theta_n, \delta_i) = 1 / [1 + \exp(\theta_n - \delta_i)] \tag{2.3}$$

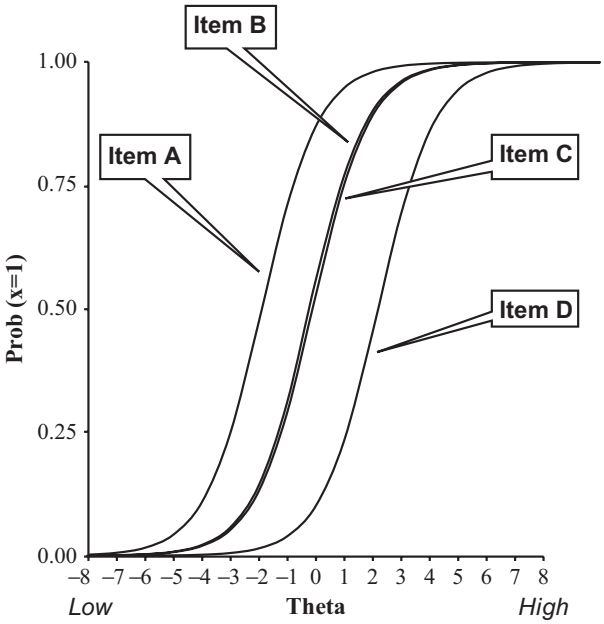


Fig. 2.2 Item response functions Rasch measurement theory (Stouffer-Toby data)

where x_{ni} is the observed response from person n on item i (0 = incorrect, 1 = correct), θ_n is the location of person n on the latent variable, and δ_i is the difficulty of item i on the same scale. Once estimates of θ_n and δ_i are available, then the probability of each item response and item response pattern can be calculated based on the model. Figure 2.2 shows four item response functions based on the Rasch model.

Model-data fit can then be based on the comparison between the observed and expected response patterns that is conceptually equivalent to other methods of evaluating a Guttman scale. Engelhard (2013) provides a description of several model-data fit indices that can be used with the Rasch model. Rasch measurement theory (Rasch, 1960) provides a framework for meeting these requirements when acceptable model-data fit is obtained. Bond & Fox (2015) provide an accessible introduction to Rasch measurement theory.

2.5 Comparison of Guttman and Rasch

This section focuses on a comparison of Guttman and Rasch approaches to invariant measurement. The first section summarizes the similarities and differences between the two models in terms of conceptual issues related to sample-invariant calibration of items, and item-invariant calibration of persons. The second section focuses on an empirical analysis of a classic data set originally presented by Stouffer and Toby (1951).

Table 2.1 Comparison of Guttman and Rasch on major issues related to person-invariant item calibration

Issues	Guttman	Rasch
1. Recognized importance of person-invariant calibration of items	Yes	Yes
2. Utilized latent variable concept	No	Yes
3. Transformation of percent correct	Percentile metric	Logit metric
4. Used item response function	No	Yes
5. Level of analysis	Individual level	Individual level
6. Assumed distribution of ability	None required	None required
7. Model-data fit	Data to *model*	Data to *model*
8. Requirements of model must be met to achieve invariance (not assumptions)	Yes	Yes
9. Person measurement	Simultaneous process	Simultaneous process
10. Level of measurement	Ordinal (non-parametric)	Interval (parametric)

2.5.1 Conceptual Analyses

Table 2.1 summarizes the issues related to person-invariant item calibration. In terms of these ten issues, Guttman and Rasch have common perspectives on six of the issues. Both of these measurement theorists recognized the importance of invariance, and their solutions included building measurement models at the individual level of analysis. Their methods did not require an explicit assumption regarding the shape of the distribution of the latent variable. Both Guttman and Rasch recognized the importance of developing a model with strict requirements for sound measurement, and then developing methods to test whether the requirements of their models have been met through the use of indices of model-data fit. The measurement theories of Guttman and Rasch reflect the idea of fitting data to *models* with an examination of whether or not these requirements have been met. If the requirements of their respective models are met to a certain degree of approximation by the data, then the desirable characteristics related to invariant measurement have been realized. Guttman and Rasch also recognized that person measurement and item calibration can be viewed as simultaneous processes.

Although Guttman and Rasch share comparable positions on six of the issues related to sample invariant item calibration, there are four important distinctions between the models. It appears that Guttman did not formally include the idea of a latent variable in the early developments of Guttman scaling. Also, Guttman used a percentile metric, while Rasch utilized a logistic transformation of the percentile metric to create logit scales. Guttman did not formally use item response functions, although as shown in Fig. 2.1 it is possible to view the deterministic nature of Guttman scaling from this perspective. Since Guttman did not use item response functions, there is no formal probabilistic model underlying Guttman scaling; Rasch measurement theory is explicitly built on a probabilistic model that provides a framework for understanding response patterns of individuals to a set of items. The

Table 2.2 Comparison of Guttman and Rasch on major issues related to item-invariant person measurement

Issues	Guttman	Rasch
1. Recognized importance of item-invariant measurement of individuals	Yes	Yes
2. Utilized latent variable concept	No	Yes
3. Avoided use of raw scores	No	Yes
4. Used person response functions	No	Yes
5. Level of analysis	Individual level	Individual level
6. Assumed distribution of ability	None required	None required
7. Used formal probabilistic model	No	Yes
8. Model-data fit	Data to *model*	Data to *model*
9. Flagged inconsistent person response patterns	Yes	Yes
10. Focused on errors in person response patterns	Major focus of model	Model-data fit essential
11. Item calibration	Simultaneous process	Simultaneous process
12. Level of measurement	Ordinal (non-parametric)	Interval (parametric)

final distinction between the two models is relates to the level of measurement. Guttman created a non-parametric approach that yields ordinal level measures, while Rasch created a parametric model that yields interval level measures.

Turning now to item-invariant person measurement, Table 2.2 summarizes the perspectives of Guttman and Rasch in terms of 12 issues. There are seven points of agreement. These points of agreement include the recognition of item-invariant measurement as a significant problem in measurement and scaling theory. Both of these theorists focus on the development of individual level models with no assumed distribution of ability. As pointed out earlier, both Guttman and Rasch approached model-data fit in terms of fitting data to *models* in order to meet the requirements of their models and yielding desirable measurement characteristics related to invariant measurement. As a part of the examination of the response patterns related to item-invariant person measurement, they also flagged inconsistent person response patterns for further study. In fact, a focus on errors in person response patterns can be viewed as a major focus of Guttman scaling, while Rasch recognized this issue more generally as a part of his concern with model-data fit. Guttman and Rasch viewed person measurement and item calibration as a simultaneous process.

Guttman and Rasch have different perspectives on five of the issues summarized in Table 2.2. As pointed out earlier, Guttman did not use the concept of a latent variable or construct. This also implies that he did not use person response functions or a formal probabilistic model for person measurement. Finally, Guttman scaling is essentially a non-parametric model that yields ordinal person measures, while Rasch measurement reflects a parametric models with the potential to create interval level measures on a logit scale.

Table 2.3 Response patterns for empirical example (Stouffer & Toby, 1951)

Person scores	Expected Guttman pattern	Observed item pattern (ABCD)	Freq	Errors	Error freq
4	1111	1111	20	0	0
3	1110	1110	38	0	0
		1101	9	2	18
		1011	6	2	12
		0111	2	2	4
2	1100	1100	24	0	0
		1010	25	2	50
		0110	7	2	14
		1001	4	2	8
		0101	2	2	4
		0011	1	4	4
1	1000	1000	23	0	0
		0100	6	2	12
		0010	6	2	12
		0001	1	2	2
0	0000	0000	42	0	0
		$k = 4$	$n = 216$	24	140

Note: Higher person scores indicate a more particularistic response, while lower person scores indicate a more universalistic response. Items are ordered from easy (Item A) to hard (Item D)

2.5.2 Empirical Analyses

Data from Stouffer and Toby (1951) are used to illustrate the indices for evaluating a Guttman scale and the Rasch model. Table 2.3 provides the response patterns for 216 persons responding to four items (A, B, C, and D). The four items in the Stouffer-Toby data have the following *p*-values (Item A to Item D): .21, .49, .50, and .69. Item A is the hardest to endorse, while Item D is the easiest to endorse for these persons. Based on this difficulty ordering, the expected patterns for an ideal Guttman scale are shown in column two of Table 2.3. Column three presents the observed patterns, and their assignment to a particular expected item pattern based on the sum scores. For example, the observed pattern [1110] sums to person score 3, and it is assigned to the expected item pattern of [1110]; there were 38 persons with this observed pattern and there are no errors. In contrast, the observed pattern [1101] also sums to person score 3, but when the expected and observed response patterns are compared, there are 2 errors. The reproducibility coefficient for these data is .84 (1−[140/(216*4)]). This value is lower than the value of .90 recommended by Guttman for an acceptable scale.

Turning now to the Rasch analyses of the Stouffer-Toby data, the FACETS computer program was used to estimate the parameters of the Rasch model (Linacre, 1989). The Wright Map showing the location of the four items and 216 persons on the logit scale is presented in Fig. 2.3. There are several features of this data that

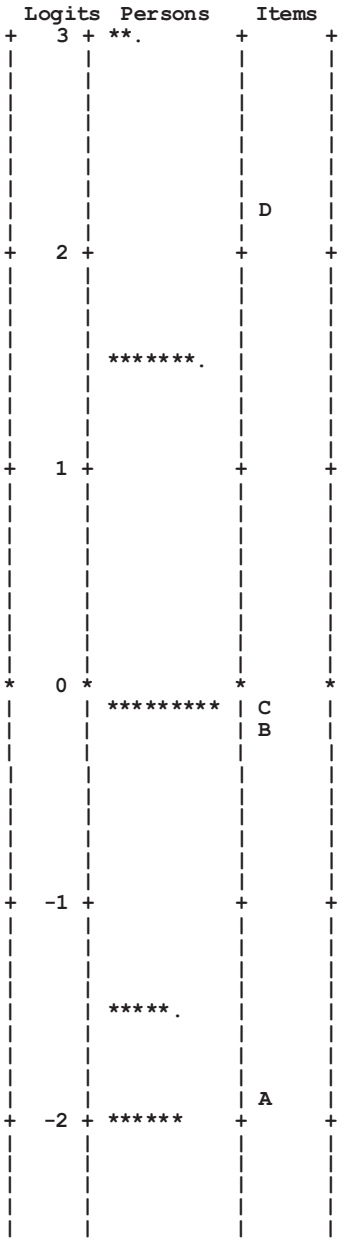


Fig. 2.3 Wright Map for Rasch Measurement Model. Note. "*" represents seven persons, and "." represents one person

Table 2.4 Summary item statistics for Rasch analyses

	Persons	Items
Measures		
Mean	.16	.00
SD	1.17	1.46
N	216	4
Infit		
Mean	1.0	1.0
SD	.7	.1
Outfit		
Mean	1.0	1.0
SD	1.4	.1
Reliability of separation	.42	.98
Chi-square statistic	283.6*	155.1*
Degrees of freedom	215	3

*$p < .01$

should be noted based on examining the variable map. First of all, there are only 4 items with items B and C having very similar locations on the latent variable. Next, the use of a four-item scale yields only 5 score groups (0, 1, 2, 3, and 4). This is reflected in the clumping of persons in five groups on the Wright Map.

Summary statistics for the Rasch analyses are presented in Table 2.4. The items are centered at .00 logits (SD = 1.46), while the person logits have a mean of .16 logits (SD = 1.17). The estimated Rasch item difficulties are −1.89, −.20, −.10, and 2.20 logits for items A to D respectively. The item response functions for these four items are shown in Fig. 2.2. The major indicators of model-data fit are based the mean square error statistics that are called Infit and Outfit in Rasch measurement theory. The Infit statistic is an information weighted mean square error statistic, while the Outfit statistic is the usual unweighted mean square statistics. See Wright & Masters (1982) for a detailed description of how these model-data fit statistics are calculated. The rules of thumb used in this paper are based on judging as acceptable any Infit or Outfit statistics between .80 and 1.20 with an expected value of 1.00. Based on this criterion, the items have good fit, while the persons have slightly more variation than would be expected by chance.

2.6 Discussion and Summary

Guttman scaling is important because it lays out in a very obvious way many of the issues and requirements that are necessary for the development of a scale that meets many of the requirements for invariant measurement. Guttman preferred to limit his approach to scaling to ranks and ordinal-level person measures that reflect a deterministic and non-parametric approach to scaling. Even though Guttman's

requirements for an ideal scale are embedded within a deterministic framework, Andrich (1985) has shown how a probabilistic model based on Rasch measurement theory (Rasch, 1960) can achieve these theory-based requirements. Andrich (1985) has pointed out the close connections between Guttman's deterministic model and Rasch's probabilistic model. In his words,

> … technical parallels between [Rasch measurement theory] and the Guttman scale are not a coincidence. The connections arise from the same essential conditions required in both, including the requirement of invariance of scale and location values with respect to each other (Andrich, 1988, p. 40).

Rasch measurement theory has become one of the most widely used item response theory models in a variety of situations (Bond & Fox, 2015). When good model-data fit is achieved, then invariant measurement provides an elegant and powerful conceptualization for guiding measurement in the social and behavioral sciences.

2.7 Personal Endnote

In his feedback on one of my seminar papers (later published as Engelhard, 1984), Ben wrote the following comment: "Do you want to take this further George? It is a fundamental integration of history and ideas" (Ben Wright, personal communication, 1980). My answer to Ben was an emphatic YES…

References

Andrich, D. A. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco: Jossey-Bass.

Andrich, D. A. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage Publications, Inc..

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Choppin, B. (1968). Item banking using sample-free calibration. *Nature, 219*(5156), 870–872.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Engelhard, G. (1984). Thorndike, Thurstone and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement, 8*, 21–38.

Engelhard, G. (2005). Guttman scaling. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 2, pp. 167–174). San Diego, CA: Academic Press (Elsevier Science).

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken [Special issue]. Measurement: Interdisciplinary Research and Perspectives (6), 1–35.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge/Psychology Press/Taylor & Francis.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (Vol. IV, pp. 60–90). Princeton: Princeton University Press.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 61, no. 4.

Loevinger, J. (1948). The technic of homogeneous test compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*, 507–530.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72*, 143–1555.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/Berlin: De Gruyter.

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review, 47*, 355–366.

Mosier, C. I. (1941). Psychophysics and mental test theory II: The constant process. *Psychological Review, 48*, 235–249.

Nozick, R. (2001). *Invariances: The structure of the objective world*. Cambridge, MA: The Belknap Press of Harvard University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Rasch, G. (1961). On general laws and meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley Symposium on mathematical statistics and probability*. Berkeley: University of California Press.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*(Part 1), 49–57.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Stouffer, S. A., & Toby, J. (1951). Role conflict and personality. *The American Journal of Sociology, 56*, 395–406.

van der Linden, W. J. (2016). *Handbook of item response theory: volume two, statistical tools*. Boca Raton, FL: CRC Press.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems*. Princeton, NJ: Educational Testing Service.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45. 52.

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

Chapter 3
Isn't Science Wonderful?

Geoff Masters

Abstract One morning in Chicago I arrived at my desk to find a note left by Ben Wright. In handwriting that filled most of the page Ben had written, 'G. Isn't science wonderful? B′. This is the story of those days, and where they led.

3.1 Daily Exhilarations

I don't remember now what excited Ben that morning he left me the note with the "wonderful science" question. He often took home what we were working on and brought it back next morning covered with ideas. Almost forty years later I still have that note—a reminder of the daily exhilaration of working with Ben as we pored over analyses of data sets, worked on the mathematics of measurement, and experimented with more succinct ways of expressing and explaining our work.

I learnt a great deal from Ben. He gave me ways of thinking and writing and a passion for discovery that have stayed with me through my career. In a general sense, what Ben and I were attempting to do was to construct deeper meaning from the specifics of experience.

In particular, what we were working on was the construction and measurement of constructs of the kind that are important in education and psychology, such as reading ability, creativity, compassion and fear of crime. The kinds of questions we were asking were: Is it a useful idea to imagine that people differ in their creativity? If so, what could be looked for and used as indicators of more creative thinking and behaviour? Is it possible to develop a numerical measure of a person's creativity? Across what range of contexts might such measures be useful? Or is creative behaviour so context specific that it is not possible to develop meaningful measures of people's creativity?

The history of educational and psychological measurement had been largely a history of people constructing instruments in the form of individual tests and questionnaires. Once constructed, these instruments often were named after the

G. Masters (✉)
ACER, Melbourne, VIC, Australia
e-mail: geoff.masters@acer.edu.au

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_3

people who developed them, such as the Stanford–Binet, Woodcock–Johnson, Otis–Lennon and Kaufman tests. A problem with this approach was that each instrument tended to be independent of all others. A parallel would have been every set of bathroom scales having a unique name and measuring in its own unit of weight. Worse, results on most educational and psychological tests were not reported in units of measurement at all; the best that could be done was to report the percentage of some population achieving a given score on a given instrument.

Ben and I were not developing instruments. Our focus was on the constructs that instruments were designed to measure. Again using a parallel from physical measurement, we were not making instruments for measuring height, we were focused on understanding height as a construct (or 'variable') and developing measurement scales marked out in units similar to centimetres and inches. As Ben often pointed out, although instruments are crucial for measurement, they should also be transient and interchangeable. We were looking beyond the specifics of instruments to the generality of constructs.

Of course in any measurement activity, interest is always in the construct, not in the instrument itself. When students take a reading test, the particular reading passages and questions on that test are not important in themselves. In fact, students are unlikely to encounter those passages and questions ever again. Individual test questions are important only as opportunities to collect information about what is really of interest—in this case, the underlying reading ability.

3.2 Engaging with Rasch

In the early 1960s, Ben had grasped the profound significance of the work of Danish mathematician Georg Rasch for establishing this desired relationship between the specifics of instruments and the generality of constructs. Rasch conceptualised a construct as a single continuum on which test items have locations reflecting their varying difficulties and test takers have locations reflecting their varying abilities. For any given item *i* scored right (1) or wrong (0), Rasch proposed that the probability P_{ni1} of a test taker *n* getting that item right rather than wrong should be governed by the distance between the ability β_n of the test taker and the difficulty δ_i of the item, and nothing else. More specifically, he proposed that:

$$\beta_n - \delta_i = \ln\left(P_{ni1} / P_{ni0}\right) \tag{3.1}$$

Ben's lasting contribution to psychometrics was to explore and promote the implications of this measurement model, particularly for the social sciences; to develop and program practical estimation and fit routines for the model; and to demonstrate the model's useful application across a wide range of measurement problems and contexts.

Beginning from my time as a graduate student at Chicago in 1977, Ben and I explored the application of Rasch's basic model for 0/1 dichotomies to items with sequences of ordered response categories. When responses to an item i are recorded in more than two categories (0, 1, … m_i), we proposed that the probability of a test taker n responding in category k rather than $k-1$ should be governed by the distance between the ability β_n of the test taker and a difficulty parameter δ_{ik} associated with the transition from category $k-1$ to category k of item i, and nothing else (Masters, 1982):

$$\beta_n - \delta_{ik} = \ln\left(P_{nik} / P_{nik-1}\right) \tag{3.2}$$

The significance of Rasch's model, whether applied to two or more response categories, was that it contained the possibility of a scale with a consistent unit of measurement and ability measures freed of the particulars of the items used (because the process took into account the estimated difficulties of the items taken).

In other words, Rasch's model introduced the possibility of relegating educational and psychological instruments and the items that make them up to their proper place in a measurement process—as interchangeable devices for making relevant observations. The model provided the basis for using item-specific observations to infer item-independent locations on an underlying construct.

Ben had another very important insight. He saw the possibility of using Rasch's model to develop deeper understandings of constructs themselves. When the difficulties of items are estimated and arrayed along a continuum, the substantive nature of the construct begins to emerge. It becomes possible to see what lower and higher levels of the construct look like. The greater the number of items located along a continuum, the richer the available information and the greater the possibility of generalising beyond the specifics of those items. And from empirically-based progressions it is often possible to develop theoretical understandings of the nature of development on the variable being measured.

The development of substantive interpretations of measurement variables was another example of our work to construct deeper meaning from specific observations. Individual test items were not the construct, but provided the only windows we had into constructs. They were concrete but incomplete manifestations—samples and examples—of the construct. The challenge was to develop from these specific examples a more generalised understanding and description of the construct itself. Much psychometric work at that time, including Rasch's seminal analyses, made little attempt to interpret measurement variables substantively. Ben's work was at the forefront of this effort.

So this is what Ben and I worked on together, starting with instruments that used ordered response categories, such as attitude questionnaires with the alternatives *Strongly Disagree*, *Disagree*, *Agree*, *Strongly Agree*; assessments of young children's psychomotor development that rated their performances on assigned tasks; and tests of problem solving that awarded partial credit for the partial solution of problems. We analysed records of responses and performances of these kinds, constructed measurement scales for the underlying constructs, and worked to interpret, describe and illustrate development along these scales.

3.3 Key Measurement Ideas

Ben gave me a way of thinking about measurement, but in truth, this was how I always thought about measurement—at least until I encountered the peculiar world of educational and psychological testing where each instrument was often a world unto itself with its own score scale and no unit of measurement. Instead, performances on tests were reported as *z*-scores, reading ages, percentiles and stanines, all designed to communicate how test takers had performed in relation to some specified population of test takers. In the 1920s L.L. Thurstone had observed that 'the very idea of measurement implies a linear continuum'. He described the fundamental requirements for measurement, including a consistent unit and the ability to make measures that are independent of the particulars of the instruments used (Thurstone, 1928). It was this understanding of measurement that Rasch's model embodied and that we were pursuing.

Other ideas also underpinned our work. One was the primacy of individuals. Rasch had appreciated the importance of focusing on individuals rather than populations. Our primary purpose was to measure individuals and then, if it was of interest, to report summary statistics for groups—rather than modelling directly the performances of populations. This focus on individuals also meant that our interest was not so much in the study of test items (which we saw as instrumental and expendable), as in understanding the performances and diagnosing the response patterns of test takers. Ben (Wright, Mead, & Ludlow, 1980) invented his 'kidmap' display for this purpose.

A second idea was that constructs are relevant across a wide range of learning and development. For example, reading ability develops over a number of years. Like Rasch, we imagined that levels of reading ability could be measured independently of age or stage of school and so sought measures that were not linked to readers' ages, but instead indicated the points students had reached on a continuum of long-term reading development. We recognised that, for some constructs, development is potentially ongoing and lifelong.

Third was the idea that the purpose of measuring is to provide information that can be used for decision making. In measuring we were doing more than ranking. The purpose of constructing substantively interpreted measurement scales was to give meaning to measures to guide action—for example, to identify starting points for treatments or interventions. Decision making was also our reason for studying the details of individuals' response patterns. And we saw measures of change over time as essential for evaluating the effectiveness of programs and interventions and for studying factors that influence change.

3.4 Learning, Curricula, and Assessments

It was only some years later, in the context of my research in school education, that I realised that my early work with Ben had given me much more than a way of thinking about measurement. It had given me a way of thinking about learning,

about learners, about teaching, about assessment, about reporting, and about the school curriculum itself. Perhaps the best way to explain this is by first outlining the traditional conception of schooling.

The starting point in the traditional conception of schooling is the school curriculum. For each grade of school, the curriculum spells out a body of content that teachers are expected to teach and students are expected to learn. This includes the knowledge, skills and understandings that students are to develop in each year of school.

The role of teachers is to deliver the curriculum for the grade—to make it engaging, interesting and relevant to students and to ensure that every student is exposed to, and has an opportunity to learn, the material that the curriculum specifies.

The task of students is to learn what teachers teach. It is accepted that not all students will learn equally well. Some more able ('more academically inclined') students will learn most of what teachers teach; others will learn much less. Underpinning this acceptance is a belief that students vary in their ability to learn what schools have to teach.

The role of assessment is to determine how well students have learnt what they have been taught. This can be done at the end of a course or part-way through a course to establish how well students have learnt to that point—information that can be helpful in identifying gaps in learning and material that needs to be retaught.

Reporting is then the process of communicating how well students have learnt what they have been taught. A student who scores 95% on a test has presumably learnt almost all of what was taught; a student who scores below 50 has presumably learnt less than half and may be considered to have 'failed'. Students' performances also are commonly reported using A to E grades.

This traditional conception of schooling is probably held by most of the community and may be appropriate if all students in the same grade of school were at more or less the same points in their learning. However, at the start of each school year, the most advanced ten percent of students in any grade are typically five to six years ahead of the least advanced ten percent of students.[1] If school were a running race, rather than commencing on the same starting line, students would be widely spread out on the running track. Nevertheless, they are judged and graded against the same finish line—the curriculum expectations for their grade.

The consequences are predictable. Students toward the rear of the pack who are two or three years behind most students and 2 or 3 years behind grade expectations struggle and, on average, perform poorly. Because the best predictor of performance in the later grades of school is performance in the earlier grades, many students perform poorly year after year. A student who receives a grade of, say D, year after year could be excused for concluding that they are making no progress at all. And worse, this form of reporting often sends a message that there is something stable about a student's ability to learn: they are a 'D student'. Not surprisingly, many less advanced students eventually disengage from the schooling process.

[1] Based on Australian data for reading and mathematics.

At the front of the pack there is a different problem. These more advanced students begin the school year on track to achieve high grades on what for many are the middling expectations for their year level. These students are not always adequately challenged and extended. Some achieve high grades with minimal effort, make relatively little year-on-year progress, and conclude that they can succeed at school without really trying.[2]

Under this traditional approach to schooling there are problems at both ends of the continuum; a proportion of each cohort falls behind and concludes that they are poor learners and some more advanced students are not challenged to achieve their potential. I believe the alternative lies in a different conception of the schooling process—one that is deeply informed by measurement and by the ideas that Ben Wright advanced.

The starting point is a different conception of the curriculum. Rather than specifying a body of content to be delivered to all students in a particular grade, this approach conceptualises the curriculum as a long-term roadmap across the years of school. This is important because teachers typically have students working at a wide spread of locations along this roadmap. Teachers of a given grade (if grades are to be retained at all) require the same deep understandings of this roadmap as teachers of the prior and subsequent grades. The roadmap itself is informed by research into the nature of long-term progress in the learning area, including research into learning sequences and progressions, the role of prerequisites, and common misunderstandings and errors that can form obstacles to student progress.

What it means to learn successfully is defined not as performance against age/grade expectations, but in terms of the progress that individuals make. Under this approach, every student is expected to make excellent progress in their learning (for example, over the course of a school year) regardless of their starting point—even the more advanced students. And the progress that less advanced students make is recognised and celebrated even if they are still some distance from achieving a notional expectation for their age or stage of schooling. Under this approach, common grade-based expectations are less important than setting high expectations for every learner's *growth*.

This approach also involves a different conception of learners. Rather than assuming that there are inherently better and worse learners as confirmed by their performances on age/grade curriculum expectations, this approach assumes that every learner is at some point in their learning and is capable of further progress if they can be engaged, motivated to make the required effort, and provided with well-targeted teaching and learning opportunities. This approach does not assume that all students should be at the same point in their learning at the same time and, consistent with more recent understandings of learning, does not put limits on what individuals may be able to achieve given the right conditions and enough time.

There are also implications for teaching. Rather than seeing teaching as the process of delivering a specified body of content to all students in the same grade of school, this approach sees teaching as a process of understanding where individuals

[2] See Griffin, P. www.smh.com.au/national/education/results-flatline-for-top-students-20130109-2cgud.html

are in their long-term learning and designing teaching and learning opportunities to meet learners at their points of need. The differentiation of a teacher's efforts in this way is much more difficult, but also more effective, than assuming that all students are equally ready for the same learning experiences.

The purpose of assessment under this approach is to establish and understand where students are in their learning. This can be done in varying degrees of diagnostic detail. When teachers establish the points students have reached in their learning they can use this information to identify starting points and to set appropriate targets for further learning, to monitor past learning progress, and to evaluate the effectiveness of their teaching strategies. Interestingly, when the purpose of assessment is to understand where students are in their learning, many traditional distinctions—such as the formative/summative distinction—become less useful.

Finally, there are implications for reporting. Current approaches to reporting performance at school are a little reminiscent of production lines in which outputs (such as agricultural produce or industrial products) are graded for their quality. Each year students are delivered the curriculum for their year level, have their performances on that curriculum assessed and graded, and then move to the next year where the process is repeated. Typically, each year is treated as a fresh start. The alternative to this kind of grading is to report the points individuals have reached in their long-term learning—indicating what they know, understand and can do at the time of assessment—as well as the progress they make over time. Information of this kind provides a basis for evaluating a student's learning and for student–teacher–parent conversations about next steps in teaching and learning.

3.5 Breakthrough

Prior to going to Chicago I had taken courses in traditional test theory. It seems to me now that traditional test theory had largely given up on Thurstone's vision of measurement. The traditional textbooks I studied began by defining measurement as 'assigning numbers according to rules', including by defining ranking as 'ordinal' measurement. They then moved quickly to introduce the normal distribution as a central element of test theory and the scaling of test scores.

I was introduced to true-scores, *z*-scores, *T*-scores, percentiles, stanines, tetrachoric correlations, Kuder–Richardson 20, and Cronbach alpha. I learnt that item difficulties could be expressed as p-values, defined as the percentage of some group answering an item correctly. It is true that, during the first half of the twentieth century, traditional test theory had developed an impressive assortment of concepts, statistics and terms, but these were not well aligned with the kind of measurement I had encountered in my undergraduate study of chemistry and physics.

Ben helped me see that the focus of traditional test theory was generally on individual instruments rather than underlying constructs. What was missing from traditional test theory was a credible method for constructing a measurement variable that would make tests mere instruments for gathering observations that could be

used to estimate locations on that variable. Rasch's model made it unnecessary to begin with assumptions about normal distributions and, with the help of Ben's XOMAT mnemonic (representing the sequence: experience, observation, measurement, analysis, theory), I could see that traditional test theory sometimes muddied the boundary between the measurement of a single variable and the analysis of relationships between different variables.

The breakthrough that occurred in educational and psychological measurement in the second half of the twentieth century required a change in mindset. Ben Wright was a leader in changing that mindset.

Upon leaving Chicago I turned my attention to how we assess and communicate learning in schools. Again, there is a traditional approach. Textbooks on assessment begin by asserting that there are multiple purposes of assessment in education and then proceed to describe different kinds of assessment, often in the form of dichotomies: formative versus summative; assessment *of* learning versus assessment *for* learning; criterion- or standards-referenced versus norm-referenced; school-based versus external; qualitative versus quantitative; continuous versus terminal. The academics who invented these distinctions often form camps advocating one kind of assessment over another, convene their own meetings and publish their own journals. Although there is an impressive assortment of assessment concepts and terms, the field is fragmented. And when it comes to communicating learning success, resort is made to marks, percentages and letter grades tied to particular courses or years of school.

What is missing from school assessment is an attempt to establish and communicate where students are in their long-term learning progress—in my view, the sole purpose of assessment in education. A breakthrough in the assessment of school learning requires a unifying change in mindset. This, in turn, requires well-constructed maps of what long-term progress in an area of learning looks like, built from research into learning sequences and based on theoretical understandings of the nature of progress—exactly the kind of map that modern measurement methods are designed to provide, that Ben Wright worked hard to promote, and that regularly aroused his excitement.

References

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554.

Wright, B. D., Mead, R. J., & Ludlow, L. H. (1980). KIDMAP: person-by-item interaction mapping (Tech. Rep. No. MESA Memorandum #29). Chicago: MESA Press. Retrieved from http://www.rasch.org/memo29.pdf.

Chapter 4
Ben Wright: A Multi-facet Analysis

Mary E. Lunz and John A. Stahl

Abstract Dr. Benjamin D. Wright believed and taught that to understand the ways of the world, it is necessary to measure all relevant aspects on the same scale. When measurements are on the same scale, it is possible to do accurate comparisons and proper ordering. In this light the multi-facet model was developed. This is a "not so scientific" study of the multi-faceted aspects of Dr. Benjamin D. Wright. Three attributes were identified for the purposes of this study: (1) Contributions to Objective Measurement, (2) Attributes as a Teacher and Professor, and (3) Personal Attributes. Data were collected and analyzed using the multi-facet model, yielding a complex pattern of results for a multi-faceted person. The real story is the development the multi-facet model. We are grateful to Ben Wright and Mike Linacre for making this tool available to measurement professionals.

4.1 Need for the Multi-facet Model

The need for the multi-facet model arose from a practical examination which was being administered by the Board of Registry of the American Society for Clinical Pathology. A practical examination was given which required subject matter experts to grade work samples from candidates. The work samples were specifically defined so that all practical examinations covered the same material. The constraints of the situation were such that the examinations could be scored by only one grader. Thus, the goal was to train the graders so that they would make similar judgments concerning candidate performance.

To accomplish this goal, a training session was provided. The first step was for each of the graders to grade the same examination. Of course, there were significant differences in their grading of the examination. This was followed by a training session in which each of the work samples was discussed along with appropriate grades for different types of performance and oral discussion of why the grades were appropriate. After the training session, the practical examination of another

M.E. Lunz (✉) • J.A. Stahl
Pearson VUE, Chicago, IL, USA
e-mail: maryhouston505@yahoo.com; john.stahl@pearson.com

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_4

candidate was graded by all of the graders. Review of the results showed that graders were still significantly different in the way they graded. This suggested that training was not sufficient to cause graders to grade similarly. The logical alternative was to account for differences in grader grading patterns so that all candidates would have the same opportunity to pass, regardless of the grader.

4.2 Beginnings

We had been working with the Rasch model on the other Board of Registry examinations. We explored ways to apply this model to account for grader differences. The analysis program of that time was BigSteps, but this program could only account for two of the three facets of the examination at one time. We could analyze graders as one facet but the candidates and the tasks would have to be combined as the second facet. Conversely we could examine the candidates as a separate facet but the graders and the tasks would be confounded. Lastly we could examine the tasks as the single facet but that left the graders and the candidates inexplicably combined, We could not put them all together in one analysis as separate elements. The dilemma was presented to Ben Wright, who had a student named J. Michael Linacre.

Mike Linacre started working on the project and developed the first version of the multi facet model in 1988 (Linacre, 1988), which accounted for all facets of the practical examination at the same time. The multi-facet model bears an algebraic resemblance to Gerhard Fischer's earlier linear-logistic test model (LLTM) (Fischer, 1973; Forsyth, Sarsangjan, & Gilmer, 1981). Surprisingly, from a user point of view, the hardest problem to overcome was the organization of the data so that it could be read into the analysis program in a format that would calibrate it all together. The development of the FACETS program along with the data organizing program FACFORM, allowed all facets of the examination to be considered simultaneously. Thus, the multi-facet model began.

4.3 Multi-facet Rasch Model Development

The multi-facet Rasch model (Linacre, 1989) was developed for the purpose of accounting for differences in grader severity, based on the premise that grader severity, developed from life-long experience and expectations, is unlikely to change substantially due to training. Along with the ability of the candidate other facets, such as the difficulty of items and tasks, can also be included in the equation. Using the multi-facet Rasch model (Linacre, 1989; Rasch, 1960; Wright, 1968; Wright & Stone, 1979) each element of each facet of a judge mediated examination is calibrated and placed on a log linear (logit) scale. Since candidates have different graders, it is advantageous to take these differences into account and determine the overall difficulty of the performance examination taken by each candidate. For example, one candidate may encounter severe graders, while another candidate gets more lenient graders, and a third

candidate gets moderately severe graders. Accounting for differences in the particular examination facets elements is critical for validity, reliability, reproducibility, and fairness.

Calibrations from the multi-facet model are estimated from the relevant observed ratings given to the candidates by graders. Grader severity estimates are calibrated from all ratings given by that particular grader to all candidates encountered during the grading session. The calibration of the severity of each grader is calculated independently and based solely on the rating given by that grader. The difficulty of the items and tasks are also calculated independently from the relevant observed ratings given to that item or task by all graders for all candidates during the entire grading.

The multi-facet Rasch model is an extension of the Rasch model and analyzes all facets of an oral examination and estimates the probability of candidate n with ability B_n achieving rating score of x on protocol i with difficulty D_i from grader j with severity C_j on task skill k with difficulty H_k:

$$\log\left(P_{nijkx} / P_{nijkx-1}\right) = \left(B_n - D_i - C_j - H_k - F_x\right) \tag{4.1}$$

B_n = ability of candidate n, the candidate facet
D_i = difficulty of item i, the item facet
C_j = severity of grader j, the grader facet
H_k = difficulty of task k, the task facet
F_x = Rasch–Andrich threshold or step calibration.

The number of facets in an examination will vary. The probability equation parameters are estimated on a natural log scale in log-odds units or logits. Each facet is calibrated independently and located on the same scale so that facets can be compared. Candidates are rated by a subset of graders. Graders rate a random sample of candidates. Candidates are often scored on the same items and tasks.

4.4 Early Research

Some of the earliest research using the multi-facet model was done with the ASCP practical examination. This practical examination had facets for candidates, graders, items, and tasks. We now had a method of analyzing the data, but had to learn how to interpret the results and at the same time contribute to improving the functionality of the multi-facet model. This involved constant interaction with Ben Wright. We did the analysis, brought it down to the University of Chicago, and Ben would always find another way to look at the data and back we would go to ASCP with more analyses to complete. Through this process we learned how to interpret the data and use it to improve the fairness of the practical examination.

One of the first issues was language. The multi-facet model created a whole new language to enable discussion of the results. Ben Wright, with Mike Linacre, determined that severity referred to graders, while difficulty referred to cases, tasks, or

items, and ability referred to candidates. From the perspective of the practitioner, this made discussion with content experts much easier and provided a language for discussing multi-facet results in articles and other publications more consistent.

A second issue was how to establish a pass point. Now that all facets of the examination were on the same scale, and the differences in examination difficulty were taken into account for each candidate, the old methods of allowing graders to make decisions about whether candidates pass or fail had to be integrated with a more advanced criterion referenced standard setting method. How that pass point was determined still had to be established. Working with Ben Wright, we developed a method of establishing a pass point that incorporated the expertise of the graders connected to the information from the scaled analytic ratings. This is reported in Lunz, Stahl, and Wright (1990) and Lunz (2000).

A third issue was interpreting the "fit statistics" in the context of the multi-facet analysis. We came to understand, with the help of Ben and Mike, that the fit statistics for each facet actually have a slightly different interpretation. Generally speaking, the infit and outfit statistics indicate consistency of grading. Case infit and outfit indicate the consistency with which graders grade within cases. It is expected that the graders will recognize the difficulty of a case, and therefore will tend to grade it higher or lower. If they do not, it shows as a misfit. Grader infit and outfit indicate the consistency with which graders grade cases and candidates. This is an intra-grader assessment as opposed to the more traditional inter-grader approach. It is expected that graders will give candidates relatively comparable grades among cases, as candidates do have an ability level, and that they will generally maintain their own level of severity. If this internal consistency does not occur, it shows up as a misfit. Candidate infit and outfit indicate consistency among the candidate and case interactions, mediated by the graders' scoring. It is expected that candidates will perform better on easier cases than on harder cases. If they do not, it shows up as a misfit and may be caused by candidate ability or grader inconsistency.

A fourth issue was the scoring design. For the multi-facet analysis, all facets had to be connected on the same scale. This provided a new challenge for the administration of performance examinations. For example, putting graders in pairs caused nesting which made it impossible to determine whether the graders were more or less severe or the candidates they rated were more or less able. It was found that it was necessary for graders to rotate partners. Another common practice of the time was having graders score specified groups of cases, projects, or essays, based on the premise that they became familiar with the area they rated. This practice also caused nesting, making it impossible to determine whether the projects were more difficult or the graders were more severe. Thus, the scoring design must include the opportunity for graders to rotate partners and projects to be scored by all or most graders in order to insure that the entire examination is on the same scale. Proving the necessity of the scoring design was sometimes a challenge.

A fifth issue involved the reorientation of judge training. Up to this point the accepted procedure was to recruit a group of experts and then submit them to an extensive training exercise. The goal of that exercise was to orient the graders to the grading rubric being used and then attempt to train the judges to use the

rubric in the same way, as multiple replicates of the same judge. Using the multi-faceted Rasch model approach, the goal of the training was to have each judge use their own internal standard consistently, based on the rubric. This recognizes that each judge is a unique expert, with their own life experiences and professional qualifications, who brings a unique perspective to the judging process. The goal of the training shifted from inter-judge consistency to intra-judge consistency (Lunz & Stahl, 1992; Lunz, Stahl & Wright, 1996).

The first serious application of the multi-facets model was for the analysis of a practical examination at the ASCP. For this examination candidates provided laboratory slides in sets to detailed specifications. Graders graded the slides using specific criteria. The multi-facet model accounted for the difficulty of the laboratory slides, the severity of the graders and the difficulty of the tasks graded, so that candidates had a comparable opportunity to pass regardless of the grader who graded their project. Using features in the Facets program we were also able to examine interactions between graders and slides to identify areas where graders were unexpectedly severe or lenient. When this information was brought to the attention of individual graders they were frequently able to identify the reasons for their different severity on those particular slides.

This beginning application of the multi-faceted Rasch model was an immensely valuable contribution to the fairness of the examination scores and outcomes. Subsequently, the multi-facet techniques were applied to other types of judge-mediated performance examinations. This early work with the FACETS model spawned an exciting period of research. A sampling of the topic explored include the equating of judge mediated examinations (Lunz, Wright, Stahl & Linacre, 1989; Stahl, Lunz & Wright, 1991), examining whether graders were generally different and consistent in their level of severity (Lunz & Stahl, 1990b; Lunz, Stahl & Wright, 1991), combining traditional tests and judge mediated tests (Stahl & Lunz, 1992) and comparing the multi-facets model to other approaches in handling judge mediated tests (Stahl & Lunz, 1993).

Ben also actively encouraged us to seek ways to spread the news about this new way of analyzing judge mediated examinations. This led to a further series of publications. (Lunz & Stahl, 1990a, 1990b; Stahl & Lunz, 1992; Lunz & Stahl, 1992; Lunz & Stahl, 1993a; Lunz & Stahl, 1993b; Lunz, Stahl & Wright, 1994; Lunz, Stahl & Wright, 1996; Stahl & Lunz, 1996)

4.5 Acceptance

The multi-faceted model is now used in many different applications, such as essay grading (Myford and Engelhard, 2002), performance assessment (Myford & Wolfe, 2004), oral examinations (Lunz & Stahl, 1993a, b), rater drift (Wolfe, Moulder, & Myford, 2001), industrial performance assessments (MulQueen & Stahl, 1997). In fact, whenever, there is an analysis with more than two facets (e.g. items and candidates), the multi-facet analysis is often used. This is just one of the many contributions to the field of measurement of Benjamin D. Wright.

4.6 A Multi-facet Analysis of Benjamin D. Wright, Ph.D.

Ben Wright was the facilitator for the development of the multi-facet model, Mike Linacre was the creator, and we (Lunz and Stahl) were the god-parents. Now it only seems fair that the circle close and the multi-facet analysis be applied to Ben Wright, himself. After Ben became ill and the Celebration of the Career and Contributions of Benjamin D. Wright was organized, it seemed only appropriate to apply the theoretical concepts and programs to analyze his multi-faceted nature. Ben believed and taught that to understand the ways of the world, it is necessary to measure all aspects on the same scale. When measurements are on the same scale, it is possible to do accurate comparisons and order "things" properly.

A not-very-scientific thirty item survey was developed. Three areas of attributes were identified: (1) Contributions to Objective Measurement, (2) Attributes as a Teacher and Professor, and (3) Personal Attributes. There were ten items within each attribute. Respondents were asked to rate each item using the 6-point scale. All data were collected via email.

The rating scale categories were defined as:

5 = Exceptional—"walks on water"
4 = Very outstanding—"head and shoulders above all others"
3 = Outstanding—"definitely noteworthy"
2 = Satisfactory—"acceptable in all respects"
1 = Marginal—"occasionally makes mistakes, but is safe"
0 = Less than satisfactory —"needs work in this area"

The data were analyzed using the multi-facet model. There were four facets in the analysis. Since there was only one candidate in this analysis, all of the ratings relate to that individual.

Facet 1 was candidates. There was only one candidate, Dr. Benjamin D. Wright.

Facet 2 was respondents. There were 10 respondents in the analysis drawn from a world-wide sample to allow the results to be generalized.

Facet 3 was the difficulty of the three attributes: Contributions, Teacher, Personal.

Facet 4 was the difficulty of each item within and among the attributes.

There were ten respondents. The range of respondent severity was .44 to −.58 logits. The reliability of separation for the respondents was only .44, indicating that the respondents tended to rate the attributes with some similarity. Attributes related to contributor, teacher, and personal showed little difference in overall assessment by the respondents. However, some of the items for each attribute were easier or more difficult for respondents to endorse. Table 4.1 shows the 30 items in difficulty order and the text of the entire item. The overall reliability of separation was .86, indicating that there were some differences in the difficulty of the items to endorse. The item fit statistics suggest that there was the most disagreement among respondents concerning items 8 (contributions to factor analysis) and 11 (orderly class presentations). Table 4.2 shows the items in Attribute order with abbreviations for each item; however, the numbers are the same. Generally, the items in the Contributions Attribute were the easiest for the respondents to agree

Table 4.1 Listing of items in order of difficulty to endorse

Number item		Measure	SE	InfitMS	OutfitMS
Most difficult for respondents to endorse					
23	Ability persuade others without offending	2.6	.33	.60	.60
27	Ability tactful when others made mistakes	2.3	.32	.70	.70
8	Contributions to the use of factor-analysis	1.7	.32	1.50	1.50
25	Concern for fellow social scientists	1.5	.32	.80	.80
14	Appropriate complexity presenting concepts	1.3	.34	.70	.70
15	Lectures generally the right length	1.2	.34	.90	.80
11	Well organized orderly class presentations	1.0	.35	1.70	1.70
17	Number of formulas and images manageable	1.0	.35	.40	.40
29	Production of books, articles, etc.	.7	.35	.90	.90
3	Development and refinement of surveys	.6	.36	.80	.80
5	Work on Measurement programming	.6	.36	1.00	1.00
7	Contributions to functional assessment	.5	.38	.70	.70
16	Style of presentation was entertaining	.4	.38	1.10	1.10
21	Quality of Speech	.3	.38	.90	.90
18	Opportunities for Class discussion	.3	.40	1.00	1.10
6	Development of fit analysis	.2	.39	1.10	1.10
9	Development of multi-facet analysis	−.0	.41	1.20	1.30
13	Provocative style encourages student effort	−.2	.44	1.50	1.40
19	Classes were generally stimulating	−.2	.44	.80	.80
1	Contributions to the analysis of MCQs	−.4	.45	.80	.80
2	Development of reporting methods “kid maps”	−.4	.45	1.10	1.10
22	Forcefulness in presenting ideas, opinions	−.7	.48	1.50	1.40
24	Ability to mentor students and colleagues	−.7	.48	.90	.90
20	Info presented has practical applications	−.8	.52	.70	.70
4	Practical applications of Rasch measurement	−1.5	.63	.90	.90
26	Ability to think logically and clearly	−1.5	.63	.80	.80
28	General level of intelligence	−1.5	.63	.90	.90
12	Knowledge of the content	−2.7	1.03	1.00	1.30
10	Contributions to Rasch Theory	−2.7	1.03	1.00	1.30
30	Dedication to field of objective measurement	−2.8	1.03	1.00	1.30

Easiest for respondents to endorse

with, while the items in the Teacher Attribute were the most difficult for respondents to agree with. The items in the Personal Attribute were in the middle.

Ben Wright’s *contributions* in promoting the use of the Rasch Model theory are well known. He freely invited people to his classes just to learn and be able to use this extremely useful method of reasoning. He was extremely knowledgeable of the content. On the down side, Ben apparently did not have the ability to persuade others without offending, and certainly often did not appear to be tactful when people made mistakes. We all applaud Ben for his dedication to the field of objective measurement. In fact many of us owe the success of our careers, in large part to Ben Wright, his research, his work, his willingness to bring everyone into the “Rasch family.”

Table 4.2 Listing of items in attribute order

	Obsvd	Obsvd	Obsvd	Fair-M	Model		Infit		Outfit	
Nu Items	Score	Count	Average	Avrage	Measure	S.E.	MnSq	ZStd	MnSq	ZStd
Attribute 1: Contribution to objective measurement										
1 1 MCQ	43	10	4.3	4.29	−.46	.45	0.8	0	0.8	0
2 2 Reporting	43	10	4.3	4.29	−.46	.45	1.1	0	1.1	0
3 3 Surveys	36	10	3.6	3.57	.65	.36	0.8	0	0.8	0
4 4 Applications	47	10	4.7	4.70	−1.56	.63	0.9	0	0.9	0
5 5 Programming	36	10	3.6	3.57	.65	.36	1.0	0	1.0	0
6 6 Fit analysis	39	10	3.9	3.88	.23	.39	1.1	0	1.1	0
7 7 Functional Assess	33	9	3.7	3.67	.52	.38	0.7	0	0.7	0
8 8 FactorAnlysis	26	10	2.6	2.56	1.75	.32	1.5	1	1.5	0
9 9 Multi-facet		41	10	4.1	4.08	−.09	.41	1.2	0	1.3
10 10 Theory	49	10	4.9	4.90	−2.78	1.03	1.0	0	1.3	0
Attribute 2: Professor and teacher										
11 11 Class	29	9	3.2	3.27	1.01	.35	1.7	1	1.7	1
12 12 Content	44	9	4.9	4.90	−2.76	1.03	1.0	0	1.3	0
13 13 Provocative	37	9	4.1	4.14	−.20	.44	1.5	0	1.4	0
14 14 Complexity	26	9	2.9	2.94	1.37	.34	0.7	0	0.7	0
15 15 Lectures	27	9	3.0	3.05	1.26	.34	0.9	0	0.8	0
16 16 Entertaining	33	9	3.7	3.71	.47	.38	1.1	0	1.1	0
17 17 Formulas	29	9	3.2	3.27	1.01	.35	0.4	−1	0.4	−1
18 18 Class discussion	34	9	3.8	3.82	.32	.40	1.0	0	1.1	0
19 19 Stimulating	37	9	4.1	4.14	−.20	.44	0.8	0	0.8	0
20 20 Applications	40	9	4.4	4.47	−.87	.52	0.7	0	0.7	0

Attribute 3: Personal										
21 21 Speech quality	38	10	3.8	3.80	.34	.38	0.9	0	0.9	0
22 22 Forcefulness	44	10	4.4	4.40	−.71	.48	1.5	0	1.4	0
23 23 Persuasive	17	10	1.7	1.68	2.63	.33	0.6	−1	0.6	−1
24 24 Mentoring	44	10	4.4	4.40	−.71	.48	0.9	0	0.9	0
25 25 Concern	28	10	2.8	2.79	1.52	.32	0.8	0	0.8	0
26 26 Logical	47	10	4.7	4.70	−1.59	.63	0.8	0	0.8	0
27 27 Tactful	20	10	2.0	1.98	2.32	.32	0.7	0	0.7	0
28 28 Intelligence	47	10	4.7	4.70	−1.59	.63	0.9	0	0.9	0
29 29 Production	35	10	3.5	3.50	.74	.35	0.9	0	0.9	0
30 30 Dedication	49	10	4.9	4.90	−2.81	1.03	1.0	0	1.3	0
Mean (Count: 30)	36.6	9.6	3.8	3.80	.00	.48	1.0	−0.1	1.0	−0.1
S.D.	8.4	0.5	0.8	0.83	1.39	.20	0.3	0.6	0.3	0.6

RMSE (Model) .52 Adj S.D. 1.29 Separation 2.50 Reliability .86
Fixed (all same) chi-square: 214.2 d.f.: 29 significance: .00
Random (normal) chi-square: 26.4 d.f.: 28 significance: .55

Regarding the *teaching* attributes, respondents agreed that Ben was a content expert; however, the complexity of his lectures was sometimes greater than desired by students. Who can forget the formulas spread over the front and side blackboards as you entered the classroom? The interesting thing is that after you heard it enough times, it all began to make sense.

Regarding *personal* attributes, Benjamin D. Wright, Ph.D. is a complex individual with many assets and liabilities. He loved his students and friends (although students perhaps did not always think so) and was willing to spend time educating individuals who wanted to apply Rasch Model theory to their practice. We all know that tact was not one of his greatest assets. He seemed to rather enjoy the role of the "measurement rogue." He taught all of us a great deal that is useful in our practice and useful in our lives. I hope that we, the next generation, are able to continue this work in theory and in practice. Of course, this was not a very scientific study, but rather a descriptive testament to a man who believed in his work and was not afraid to share his thoughts with others.

4.7 Conclusion

The multi-facet analysis project that Ben nurtured has become an accepted tool for the analysis of examinations and surveys that have more than two facets. It is useful for understanding the examination process, sorting out the factors that contribute to candidate outcomes, and understanding how the facets of the examination fit together on the same scale. We are grateful to Ben Wright and Mike Linacre for making this tool available to measurement professionals like ourselves and many others.

References

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *ACTA Psychologica, 37*, 359–374.

Forsyth, R., Sarsangjan, V., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5*, 175–186.

Linacre, J. M. (1988). *FACETS, a computer program for the analysis of multi-faceted data*. Chicago: MESA Press.

Linacre, J. M. (1989). *Multi-faceted measurement*. Chicago: MESA Press.

Lunz, M. E., Wright, B. D., Stahl, J. A., & Linacre, J. M. (1989). *Equating practical examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Lunz, M. E., & Stahl, J. A. (1990a). A comparison of intra and interjudge decision consistency using analytic and holistic scoring criteria. *Journal of Allied Health, 19*(2), 173–179.

Lunz, M. E., & Stahl, J. A. (1990b). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*(4), 425–444.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1990). *Criterion standards from benchmark performances for judge intermediated examinations*. Paper presented at the annual meeting of the American Educational research Association, Boston.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1991). *The invariance of judge severity calibrations*. Paper presented at the annual meeting of The American Educational research Association, Chicago.

Lunz, M. E., & Stahl, J. A. (1992). New ways of thinking about reliability. *Professional Education Researcher Quarterly, 13*(4), 16–18.

Lunz, M. E., & Stahl, J. A. (1993a). Impact of examiners on candidate scores: an introduction to the use of multifaceted rasch analysis for oral examinations. *Teaching and Learning in Medicine, 5*(3), 174–181.

Lunz, M. E., & Stahl, J. A. (1993b). The effect of rater severity on person ability measures: a Rasch model analysis. *American Journal of Occupational Therapy, 47*(4), 311–318.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54*(4), 913–925.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibrations. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 99–112). Norwood, New Jersey: Ablex.

Lunz, M. (2000). Setting standards on performance examinations. In M. Wilson, G. Engelhard & K. Draney (Eds.), *Objective measurement: Theory into practice, Vol. 5* (pp. 181–202). Stamford, CT: Ablex.

MulQueen, C., & John A. Stahl, (1997). *Multifaceted Rasch analysis of 360° performance assessment data*. Paper presented at the annual meeting of the Society of Industrial and Occupational Psychologists, St. Louis.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.

Myford, C. M., & Engelhard, G., Jr. (2002). Evaluating the psychometric quality of the National Board for Professional Teaching Standards Early Childhood/Generalist assessment system. *Journal of Personnel Evaluation in Education, 15*(4), 253–285.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Stahl, J. A., Lunz, M. E., & Wright, B. D. (1991). *Equating examinations that include judges*. Paper presented at the annual meeting of the American Educational research Association, Chicago.

Stahl, J. A., & Lunz, M. E. (1992). Impact of additional person performance on person, judge and item calibrations. In M. Wilson (Ed.), *Objective measurement: theory into practice* (Vol. 2, pp. 189–206). Norwood, NJ: Ablex.

Stahl, J. A., & Lunz, M. E. (1993). *A comparison of generalizability theory and multi-faceted rasch measurement*. Paper presented at the annual meeting of The American Educational Research Association, Atlanta.

Stahl, J. A., & Lunz, M. E. (1996). Judge performance reports: media and message. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 113–125). Norwood, New Jersey: Ablex.

Wolfe, E. W., Moulder, B. M., & Myford, C. M. (2001). Methods for detecting differential rater functioning over time (DRIFT). *Journal of Applied Measurement, 2*(3), 256–280.

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

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

Chapter 5
Reflections on Benjamin D. Wright: Pre- and Post-Rasch

Herb Walberg

Abstract Of the contributors to this volume, I probably knew Ben Wright among the longest. It was 1960 when I first became a University of Chicago doctoral student. In Ben's statistics class, he announced that he needed a research assistant to help analyze data. Two of us applied, and I counted myself lucky to be chosen. In those days, I had the top qualifications of having previously key-punched several IBM cards. I'll recount some of my experiences in those years, and in the years since.

5.1 Computing Semantic Differentials

In my first and subsequent experiences, Ben exuded exuberance in whatever research he was engaged. Indeed, he inspired me to spend long hours on his research. As a consequence, I neglected his course (for which I ultimately received a grade of "C") and other courses to spend perhaps 45 h a week on my "half-time" assistantship.

Ben had me using a huge vacuum-tube computer, which filled a large basement room, to analyze Semantic Differential (SD) data. This SD instrument, made popular in the 1960s by Charles Osgood, called for respondents to rate concepts such as "Kitchen Cleanser" on five or seven-point scales between such adjectives as "warm" and "cool." For market research, Ben had dozens of concepts and scads of adjectives that could be analyzed in innumerable ways, and I could hardly keep track of them. So obsessed with getting the analyses finished, I hardly knew the purpose of the research, which may have been to discover perceptions of products and what factors lead to favorable views.

To reduce the data to manageable dimensions, Ben had me run many factor analyses. A single one might take hours, and the data often had to be rerun since the computer would sometimes blow a tube halfway through. At least it didn't take the months that a previous generation of students took for hand calculation. What the students of Louis Thurstone, the University of Chicago's famous mental-abilities

H. Walberg (✉)
University of Illinois, Chicago, IL, USA
e-mail: hwalberg@uic.edu

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_5

theorist, could do by hand in 6 months took me 6 h. Now desktop computers can perform such computational feats in 6 s.

The research assistantship with Ben was exhilarating but frustrating. Though I knew the experience would change my life, I felt I had to concentrate on my doctoral studies and resigned the assistantship. I got a professorial sinecure, which tripled my income and halved my working hours so I could turn to my examinations and dissertation. (In retirement, I continue to think academic professing is far better than working for a living.)

When I was ready to start a dissertation, I reflected that Ben had previously been in physics, a field in which people tend to get their degrees early and often reach old age at 35. I guessed he might have retained the view of research as a young person's game. Being 24 years old, I thought Ben was my man, and he was; I finished, thanks to Ben, just after turning 26.

At the time of my last two years of graduate study, Ben, having been mentored by the famous psychoanalyst Bruno Bettelheim at the Orthogenic School for autistic children, was interested in applying psychoanalytic theories to the study of teaching, particularly the psychology of beginning teachers. Ben had designed an SD to explore the differences among such concepts as Myself, Myself as a Teacher, and, as I recall, My Ideal Teacher, My Father, and My Mother. Though I have long ago lost track of them, his assistants at the time were Shirley Tuska, Barbara Sherman, and Douglas Stone. I think Shirley and Barbara employed the massive (for its time) database Ben assembled, as I did; and I believe we all finished our University of Chicago doctoral studies about 1964.

5.2 Early Days with Rasch

When I finished my degree, I had worked for 12 years, beginning at age 13, and faced three or four more decades of university leisure. Before starting, I decided to seize the opportunity to circumnavigate the globe. Ben said he'd be spending time in Copenhagen and invited me to look him up, which I did. When I got there, Ben, with his usual enthusiasm, told me he was on leave to study the mathematics of testing with the mathematician Georg Rasch. He excitedly told me the great significance of Rasch's views and how they might transform the somnolent psychometrics, or measurement of knowledge, skills, and attitudes.

My great mistake was to take too lightly Ben's new enthusiasm and to continue my interests in statistical computing that I had learned from him and the substantive psychology that I had learned from other Chicago faculty—Ben Bloom, Bruno Bettelheim, Phil Jackson, Fred Lighthall, Jack Getzels, and Herb Thelen.

In the early 1960s, the quantitative faculty, later to begin the MESA (Measurement, Evaluation, and Statistical Analysis) program, was a part of educational psychology at the University of Chicago. Later Bert Masia, Darrell Bock, and David Wiley also joined the MESA group, and it became one of the best places in the world for educational, psychological, and social science measurement. There physicist Lord

Kelvin's saying, chiseled on the portal of one of the buildings, was taken seriously: "When you can measure what you are speaking about, and express it in numbers, you know something about it."

The division of substantive and methodological psychology within education exemplified rapidly increasing academic specialization during the last century, particularly in the last several decades. Psychology itself had split away from philosophy. Theoretical and empirical psychology divided as did applied fields such as clinical, educational, and industrial psychology. Even the methodological fields comprising MESA became more specialized with separate handbooks, journals, and academic societies. No one could know it all. Large-scale developmental and research efforts usually required teams of specialists.

Even though measurement was hardly my passion in the late 1960s, there was a severe staff shortage in the field. My smattering of ideas—picked up from working with pre-Rasch Ben and teaching first-year courses in testing—got me a job invitation at Educational Testing Service in Princeton, New Jersey. I supposed the opportunity came along because I had studied with Ben Bloom, author of the famous *Taxonomy of Educational Objectives*. Ben Wright, then making the transition from psychoanalysis to psychometric analysis, was still a young man, much less well known than he was later to become.

After a few months at the Testing Service, I got a manuscript from Ben, who asked me to show it to Fred Lord, who seemed the grand old man of "classical measurement," which favored the "three-parameter" model. In contrast, Ben and other Raschians favored the "single-parameter" model. I turned over the manuscript to Lord with little explanation since I lacked (and still lack) sufficient knowledge to favor either side.

I felt pangs of guilt that I could not have been a skilled diplomat to bring some understanding and mutual appreciation between the classicists and Raschians, who seemed increasingly at odds in the years to come. As Lincoln said, "A house divided against itself will not stand." In any case, it is probably just as well that I didn't try any diplomacy since my ignorance of the technicalities would have only made things worse.

A few months at Educational Testing Service convinced me that the leisure of the academic theory class would be better for me. Against all rules of decorum, I impetuously informed a beginning Harvard dean that I might be free for new employment. But it worked, and I was invited for a day of coffee and meals with him and several faculty.

Having heard nothing for a week or so, I called the dean's secretary and found that a job offer had gone to my Chicago address. Later, I learned that eminent psychometrist Phillip Rulon had retired and that Jack Carroll was leaving Harvard for North Carolina to synthesize factor analyses of mental abilities. I was being brought in to teach courses in measurement.

During these years, I realized that measurement was an adolescent, if not nascent, field that would become increasingly important not only in educational, psychological, and social science research but in government policy. In particular, policy makers and educators needed valid measures of academic learning and other indications of success to know what works. I also realized that I had rare 1960s computer experience and had worked with two eminent psychometrists—Ben Bloom and Ben Wright.

I stayed in touch with Ben Wright, who told me he was planning to visit Bruce Choppin of the United Kingdom's National Foundation for Educational Research. In 1964, the year I finished my dissertation, Bruce had begun as Ben's first "Rasch" student, and I asked to tag along on the trip to visit Bruce in England and learn about what he and Ben were doing. Ben, of course, has since educated people from all over the world in the rigorous design, analysis, redesign, and calibration of tests and test items.

In the United Kingdom, Bruce was pioneering large "banks" of test items and longitudinal testing, which would enable local educational authorities and nations to trace "the value added" progress of achievement of primary and secondary students. Though he died young and unexpectedly in 1983, Bruce contributed much to the International Association for the Evaluation of Educational Achievement and was elected president of the British Educational Research Association. He helped bring Rasch's and Ben's ideas to Europe and other continents (McArthur, Postlethwait, Purves, & Wright, 1985).

5.3 Policy Support for Advances in Measurement

My short visit with Ben and Bruce convinced me that I'd never be able to keep up. They were making fast progress not only in the United States and the United Kingdom but also in promulgating their Raschian ideas around the world. But I could be a cheerleader, and ever since, I was, as chair of the Design and Analysis Committee of the National Assessment Governing Board, chair of the Education Indicators Committee for the Organization for Economic and Cooperation Development, and a member of various advisory and governance groups.

I saw little of Ben in the 1980s; we were both busy. But I retained a policy interest in measurement since I often testified on achievement issues before Congress and federal courts, collaborated with others in designing and analyzing large-scale surveys, and advised my own and others' doctoral students on research methodology. Around 1990, however, I met William Fisher, who had been a student of Ben's long after my time. Bill brought me up to date on Raschian analysis. Thirty years too late, I realized its broader significance for education policy and practice, and it struck me as hugely important for nonspecialists in measurement to learn about. As chair of the editorial board of the *International Journal of Educational Research*, I invited Ben and Bill to compile the best work in the field for a special issue. Our editorial board was highly pleased with the special issue they submitted (Fisher & Wright, 1994), and I hope it brought Raschian insights and their significance to a large group of researchers around the world.

In this fourth phase of his career after physics, psychoanalysis, and research on teaching, Ben brought an obscure psychological measurement idea from a Danish mathematician to the United States and later to the world. In the last three decades of his career, charismatic Ben inspired others to carry out applied methodological research that has great potential for increasing the productivity of education and other fields.

More than ever, world policy makers see the causal connection of school achievement to economic progress and individual welfare. It seems an education law that what gets measured gets taught—if not learned. To hold students and educations systems accountable for meeting outcome standards, we require efficient, scientifically designed tests. My experience in the 55 years since I first met Ben lead me to believe that we are now beginning the "Golden Age of Measurement" in education, psychology, and the social sciences. Decades ago, visionaries Georg Rasch, Ben Wright, Bruce Choppin, and others showed us the way.

Can we measure up to their standards?

References

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

McArthur, D. L., Postlethwaite, T. N., Purves, A. C., & Wright, B. D. (1985). Introduction: Memories of Bruce Choppin. *Evaluation in Education, 9*(1), 5–7.

Chapter 6
Reflections: Ben Wright, Best Test Design and Knox's Cube Test

Mark H. Stone

Abstract Writing this chapter past age 80 occasions some reflections upon my life as it intersected with Ben Wright. I presented some of these thoughts at a symposium given in Ben's honor in Chicago in April of 2003. More comments can be given here to serve a wider audience. In reflecting on past times, I do so with great fondness for Ben, and in appreciation of my long friendship with him lasting more than fifty years. Not only have I had the occasion of his friendship, we shared the collegial opportunity of writing about Rasch measurement. These activities occasioned the opportunity to meet and share a friendship with many others who also acknowledge a fondness for him.

6.1 Ben Wright

I went to the University of Chicago to study music following my undergraduate work in music and psychology. I was not sure which to pursue. Should it be solely music or clinical psychology? (I still have not decided.) The University of Chicago environment cultured both areas for me without my having to make a choice. My teacher for performance studies was the University Organist, Edward Mondello, with whom I had already studied pipe organ and methods as an undergraduate. He presided at Rockefeller Chapel giving recitals each term, and teaching his students, including me.

I encountered Ben in 1955 at a seminar conducted under the rubric *Committee on Human Development.* Ben had earned his Ph.D. in this area following his move from physics to the social sciences. Ben had previously worked at the Orthogenic School under Bruno Bettelheim. My clinical work with psychotic children at Chapin Hall in northwest Chicago (now located on the campus of the U of C) had already brought me in touch with Bettelheim. Some of the young children I worked with at Chapin Hall were relocated to the Orthogenic School. In addition, I enrolled in Bettelheim's course in dream interpretation and his seminar in psychoanalysis.

M.H. Stone (✉)
Adler School of Professional Psychology, Oswego, IL, USA
e-mail: Markhstone2@sbcglobal.net

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_6

Ben was involved in these seminars, so we met again. He and Bettelheim had authored several papers together that were required reading. While Ben had completed his (Freudian) studies in psychoanalysis at the Institute of Psychoanalysis in Chicago, I did my psychoanalytic work (Adlerian) at the Adler Institute in Chicago, where I subsequently became academic dean.

After 2 years of study on campus I decided to complete my graduate work in music at the Chicago Musical College (CMC). This was not much of a change because, at that time, CMC operated a joint doctoral program with the U of C. My interest moved to music theory. I also transferred to the clinical psychology department at Northern Illinois University because the program operating under the rubric *Committee on Human Development* did not meet the requirements established by the State of Illinois to allow me to sit for my license examination in clinical psychology. Upon graduation and while beginning to practice clinical psychology I returned to the U of C to take further coursework with Beck (Rorschach), and Ben Bloom, Darrell Bock, John Bormuth, and Ben Wright in the Department of Education's Measurement, Evaluation, and Statistical Analysis (MESA) concentration.

From 1965 to 1975, I worked part-time at Social Research, Inc., (SRI) in Chicago. Ben had worked there from about 1955. SRI was an interesting company. It was a research and consulting firm founded by U of C professors—several of whom were also my teachers. In the late 1960s, Ben took the train from Hyde Park to the loop every Thursday to work at SRI. SRI was headed by Burleigh Gardner, Ph.D. Gardner was an anthropologist from Harvard who had earned his doctorate at the U of C. With Davis, Gardner, and Gardner (1941) he co-authored several books, including *Deep South*. Mary Gardner, Burleigh's wife, was a co-author also. She was an editor of medical books, but everybody was surprised to learn at her funeral she anonymously authored numerous racy, romance novels for fun and profit.

At SRI, Ben and I worked on a large number of research studies, especially for advertising and marketing firms nationwide. Lee Rainwater, Richard Coleman, and Gerald Handel (1962), all from the U of C (the latter two from the Committee on Human Development), spent time at SRI. Together they published *Workingman's Wife*, using data collected from SRI surveys. Ben's role was especially prominent in the construction of questionnaires, interview schedules, research design, and data analysis. Careful pre-planning of these instruments and studies was a hallmark of Ben's work, together with his careful attention to detail.

Ben abstracted Chapter 8 from Thurstone's *Multiple-Factor Analysis* (1947) into a handbook for us to use in doing factor analysis by hand. Prior to microcomputers we used this handbook for analyzing data from many studies. Before Rasch measurement occupied his activities it was factor analysis that was Ben's passion. Factor analysis was applied to a great amount of data from the large number of contracts SRI received from the tobacco industry for studies on packaging, use of color, and advertising effectiveness. (Interestingly, no one smoked at SRI.) This methodology and analysis were also applied to employee satisfaction studies for various companies, and scales for use in college and university alumni satisfaction studies and course evaluations. SRI also conducted numerous studies of local TV news broadcasting in cities nationwide to determine viewer

interest and to set advertising rates. Ben participated in these studies both as a psychologist and data analyst.

While Ben was the statistical consultant, he also participated in numerous activities developed at SRI. William Henry, also from the U of C, author of *The Analysis of Fantasy* (1956), developed an industrial set of Thematic Apperception (TAT-like) cards that SRI used for executive assessment. We used these cards as part of our assessment battery for evaluating executives for hiring and promotion. Ben created a similar set of TAT-like cards that we used with children. A variety of other psychological scales were produced for use in the surveys and studies conducted by SRI.

6.2 Best Test Design

Noontimes we spent discussing Rasch. Ben gave me some of Rasch's original statistical papers because I can read Norwegian (Our family name is Ohlsen-Stene). With the aid of a Danish dictionary I went through some of his papers. (Norwegian is similar in composition inasmuch as Denmark had once ruled Norway and Dano-Norwegian or Bokmål is the Norwegian literary language.)

From these discussions over lunch we moved to working out some explanatory examples to satisfy my interest and curiosity. Finally, we decided to write a book and formalize these activities. On almost every Sunday I went to the U of C campus to meet Ben at his office, or I went over to Harper Avenue to sit at the kitchen table with Ben. Following our talks, I went home to type drafts for the next meeting. This was, of course, in the years before PCs, so most of our work, and all of the composition, was analog.

After some frustrating production delays in negotiating with several publishers, we decided to have the book printed at the U of C and do all the preliminary work ourselves. Today we can do these tasks handily thanks to laptops and software, but at that time typing drafts had to be followed by typesetting, and we needed this latter service. Our plan by this time was to make the pages 8 by 11–1/2 inches to accommodate the tables and figures we thought were required for a step-by-step explanation of how to calibrate items by hand, and to present the other tables and graphs we developed for our book, which we entitled *Best Test Design* (BTD). It necessitated careful work.

I engaged a young woman recently from Germany who was working in the Loop. She had Americanized her first name to "Sam" and I brought the final drafts of the chapters to Sam to be typeset when I was working downtown at SRI. For the next half-year Ben and I worked each Sunday, and during the week I worked with Sam on getting the final pages in correct form. As more than half the chapters took shape Ben indicated that he appreciated the fine work that Sam was doing and wanted to meet her. Sam did excellent work with great attention to detail. She was married, working full-time and completing this task for us during her noon hours, after work in the evenings, and on weekends. She had no time or inclination to visit the campus. Ben finally insisted that he must at least talk with her. I gave him her phone number.

The following morning her husband came to my office at SRI. He said that his wife was finished working on the project and following her conversation with Ben she had burned the final drafts in their fireplace! Ben didn't want to talk about the phone call he had with Sam when I later broached him about this incident.

I had typed all the drafts, and the final pages were re-typed using composing machines with the final copy pasted on to 2′ by 3′ make-up boards, which were covered with tracing paper to protect the copy. Fortunately, I had kept all the earlier drafts in my basement. I still have them there, and when I die my children and grandchildren will probably ask, "Why did he keep all this junk?"

I was able to "reconstitute" everything lost in the blaze of fire resulting from Sam's apparent rage. I brought a new typed edition to Ben. Without a word from him we continued with the remaining chapters. I have never inquired of either Ben or Sam about what happened in their conversation, but anyone who knows Ben knows his (Greek/Socratic) "daemons" and some things are best left alone.

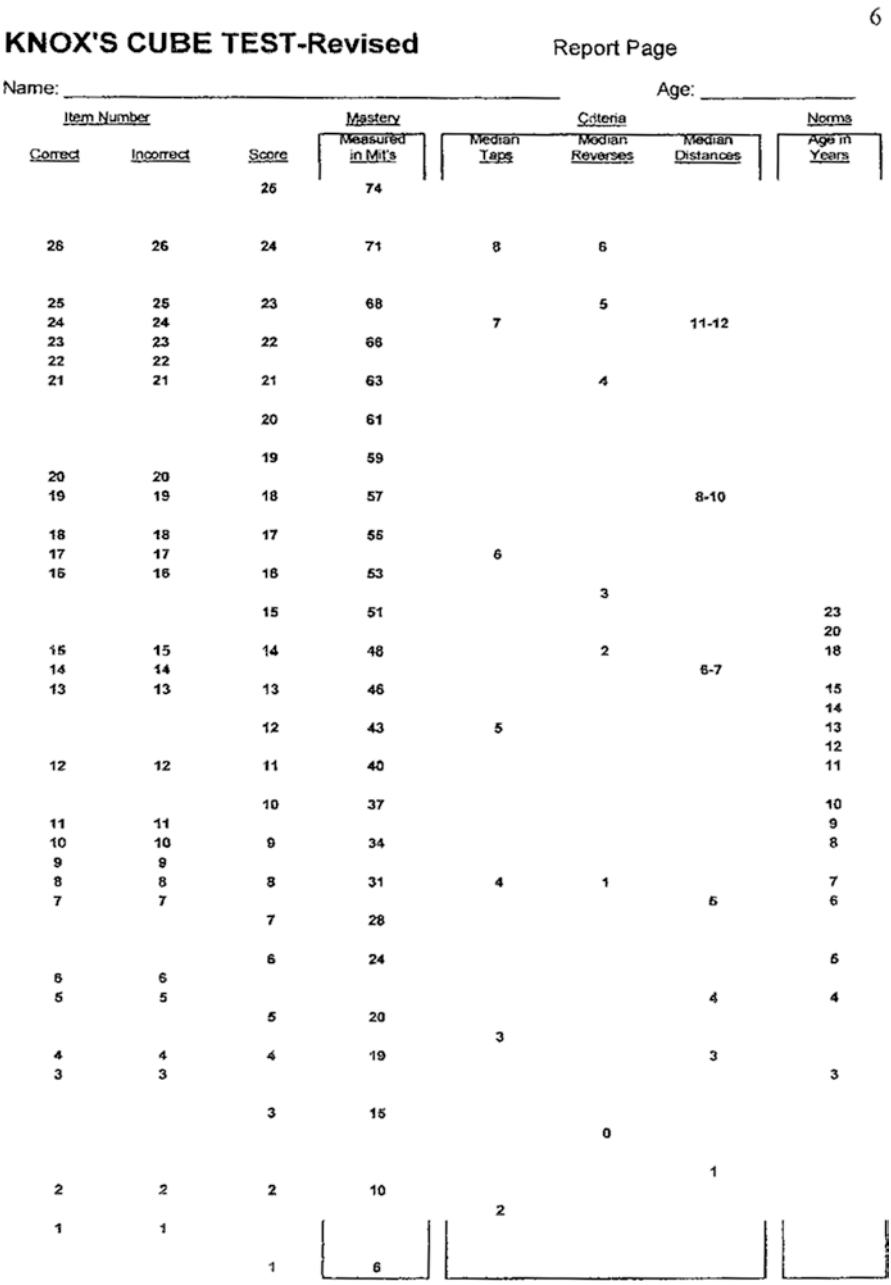
I next engaged Betty Stonecipher to complete the remaining chapters. She did excellent work also. My wife's name is Betty, and she too worked at SRI. Stonecipher seemed too coincidental to be real, and for a long time Ben wrongly surmised that I was trying to put something over on him. He asked to meet Mrs. Stonecipher, just as he had asked to meet Sam, but I adamantly refused to comply, not wishing to see another fire erupt. When BTD was finished and printed, I invited Mrs. Stonecipher to meet Ben so he could know there actually was such a person, but I am not convinced he still didn't think I engaged her as a ploy to continue what he thought was a deception.

Out of the fire and ashes arose BTD once more. I finally delivered all the plates to the U of C print shop and the book made its appearance. Several hardbound editions were printed, but the bulk of BTD that have been produced for sale have all been paperback versions. I still have some hardback copies in my library.

6.3 Knox's Cube Test

I suggested Knox's Cube Test (KCT) for a simple example to illustrate item calibration by hand. I encountered the KCT during my clinical psychology internship when learning the *Arthur Point Scale*, a test battery for evaluating the intellectual functioning of children. Rasch's approach seemed ideal for studying the results I was obtaining from children and adults. The KCT offers a simple variable, making it easy to see the relationship between items and the procedures given in BTD.

In 1915, Dr. Howard Knox, M.D., retired from the U.S. Public Health Service and returned to private practice. Prior to this date, Knox served at Ellis Island where, among other duties, he developed several instruments for assessing immigrants seeking to enter the United States. One of these tests, a cube tapping imitation, was designated Knox's Cube Test (KCT). The original cubes remain on display at the Ellis Island Museum. The test materials were subsequently published by the Stoelting Company of Chicago. Stoelting continues to market the KCT, with the most recent revision identified as KCT-R (Stone, 2002a).



6

KNOX'S CUBE TEST-Revised Report Page

Name: _____ Age: _____

Item Number			Mastery	Criteria			Norms
Correct	Incorrect	Score	Measured in Mit's	Median Taps	Median Reverses	Median Distances	Age in Years
		25	74				
26	26	24	71	8	6		
25	25	23	68		5		
24	24			7		11-12	
23	23	22	66				
22	22						
21	21	21	63		4		
		20	61				
		19	59				
20	20						
19	19	18	57			8-10	
18	18	17	55				
17	17			6			
16	16	16	53				
					3		
		15	51				23
							20
15	15	14	48		2		18
14	14					6-7	
13	13	13	46				15
							14
		12	43	5			13
							12
12	12	11	40				11
		10	37				10
11	11						9
10	10	9	34				8
9	9						
8	8	8	31	4	1		7
7	7					5	6
		7	28				
		6	24				5
6	6						
5	5					4	4
		5	20				
				3			
4	4	4	19			3	
3	3						3
		3	15				
					0		
						1	
2	2	2	10				
				2			
1	1						
		1	6				

Fig. 6.1 Knox Cube Test scoring form

The KCT (Stone & Wright, 1980) and the KCT-R (Stone, 2002a) kits consist of four one-inch black cubes. The four blocks are fastened to a 10 by 1 inch board two inches apart. A fifth black cube is used by the examiner to tap a pattern at one tap per second. The examinee is asked to imitate the pattern. A succession of increasingly difficult taps continues from two- to eight-block tapping patterns. Knox's original edition contained only a short progression of cube tapping patterns, and the blocks were arranged by hand on a table. The current revised edition, KCT-R (Stone, 2002a), contains 26 items, with 160 more available in a bank. How Knox first conceived this test is unknown, but it shows amazing ingenuity. Deceptively simple in item structure, the increasingly complex pattern of taps produces a clearly defined item difficulty sequence applicable over the life span, from ages 2 to 90.

The KCT was included in test batteries by Pintner (1915), Arthur (1947), and Babcock (1965), and then in revised forms by Stone and Wright (1980), and Stone (2002a). There were other revisions, but these made substantial deviations from Knox's original cubes in color, block spacing, block size, etc.

Use of the same basic format by Pintner, Arthur, and Babcock permitted Wright and Stone in BTD (1979) to link all these earlier versions using common items from each edition. This linking process (described in detail in BTD) facilitated a common ability scale expressed in Mastery Units (MITs) for all of these KCT editions. Examiners were thus enabled to connect any version of the KCT to the others using the absolute measures reported in MITs. Measures derived from the earlier editions of the KCT could be compared to those of a later one or vice versa. Linking of this kind offers definite advantages to a clinician now able to compare individual performances across different editions, and across a long time period. The tables given in Stone (2002a) for making such comparisons are a consequence of the procedures and analyses first explained in *Best Test Design* (Wright & Stone, 1979). Figure 6.1 shows the KCT-R scoring form.

The measures reported for KCT-R in mastery units range from 3 to 80 MITs. In addition to the number of taps, the number of reverses in direction, and the total distance in taps covered in the tapping pattern are believed to be causal determinants of observed item difficulty. Table 6.1 shows the 26 items of the KCT-R with their logit difficulties and the three characteristics. The complete KCT-R range is 12.61 logits. This is an exceptionally wide spread for item difficulties and measures from a single test. Table 6.1 from the Manual reports all the KCT-R items, from #1 at logit difficulty −6.75 (composed of 2 taps, 0 reverses and a distance of one block) to #26 at 5.86 logits (with 8 taps, 6 reverses and a combined distance of 11 blocks).

Given the difficulty values for these 26 items, and the three characteristics of each item, regression analysis suggested how these characteristics operate in producing the total range of item difficulties. Stenner and Smith (1982), using data from Stone and Wright (1980), found that 93% of the total variance in item difficulty was accounted for by taps and distance. Using data from the KCT-R, Stone (2002a) produced a regression equation with 96% of the variance accounted for by taps alone, with a beta value of 0.91. It is clear from these two studies that the number of taps dominates the production of item difficulties with distance and reverses contributing less causal influence.

Table 6.1 KCT-R item difficulty in logits with taps, reverses and distance

Item	1	2	3	4	5	6	7
Difficulty	−6.75	−6.38	−4.89	−4.86	−3.63	−3.51	−2.60
Taps	2	2	3	3	3	3	4
Reverses	0	0	0	0	0	0	0
Distance	1	3	3	4	3	4	5
Item	8	9	10	11	12	13	14
Difficulty	−2.38	−2.07	−1.96	−1.55	−0.91	0.33	0.42
Taps	4	4	4	5	5	5	5
Reverses	2	2	1	1	1	3	3
Distance	5	7	5	3	3	6	6
Item	15	16	17	18	19	20	21
Difficulty	0.71	1.77	1.97	2.09	2.74	2.95	4.22
Taps	5	6	6	6	6	6	7
Reverses	2	3	4	2	2	4	4
Distance	7	9	9	8	9	10	11
Item	22	23	24	25	26		
Difficulty	4.42	4.52	4.67	4.84	5.86		
Taps	7	7	7	7	8		
Reverses	4	3	4	5	6		
Distance	10	12	11	10	11		

What has been accomplished thus far with the KCT-R with respect to item generation? A seminal paper by Cronbach and Meehl (1955) addressed *construct validity,* a term which had earlier been introduced by the recently published *Technical Recommendations* (1954). Cronbach and Meehl explicated the implications derived from the introduction of construct validity in this handbook by writing, "A construct is some postulated attribute of people, assumed to be reflected in test performance." They define the problem as essentially, "What construct *accounts for variance* in test performance?" (p. 283, my emphasis).

Knox, much earlier, presumed that his test provided evidence of mental competency arising from an examinee's latent ability for repeating the tapping patterns presented to her or him. How this process was actually determined, and any criteria for competency Knox used, remain unknown. However, this does not compromise the insight he used to evaluate mental competency for thousands of persons. His insight is especially rich given the exceedingly large and varied population representing many languages and cultures he was required to assess. What accounts for variance within and between persons are the item characteristics of the tapping patterns and the individual differences produced by imitating these tapping patterns.

Construct validity, according to Cronbach and Meehl (1955), should take us beyond the local scores of any version of the KCT to generic measures transcending specific particulars. This was Rasch's goal (Rasch, 1980). Linking the various editions of the KCT to the absolute scale denominated in MITs provides values for eight tapping patterns common to the five editions (Table 6.2). From these values

Table 6.2 KCTR linking patterns and ensemble means

Tapping	Ensemble	Taps	Reverses	Distance
2-1-4	16.4	3	1	5
3-4-1	19.8	3	1	5
1-4-3-2	25.6	4	1	6
1-3-2-4	30.2	4	2	6
3-1-4-2	35.4	4	2	9
2-1-1-3-3	37.6	5	1	3
2-3-4-4-3	40.2	5	1	3
1-3-2-3-1	44.8	5	3	7

the Gibbs/Einstein *ensemble* for each pattern can be expressed as the mean. The idea of the ensemble can be found in *The Philosophy behind Physics* (Brody, 1993). Brody writes, "Averaging over this set of models—technically known as an ensemble—is a trivial operation mathematically, but one that has the power of creating new concepts of quite different characteristics" (p. 126).

Figure 6.2 shows the plot of the ensemble means regressed on taps. A regression line resulting from these values produces $R^2 = 0.88$, which seems remarkable in its consistency, given the chronological years covered—almost a century. Sampling variations, as well as the varying conditions from which the data were compiled over this time period, are now summarized via the ensemble approach. This result meets the expectation Brody (1993) specified, producing "higher-level model building from using these averages" (p. 126).

Regression applied to item generation has been criticized by some who contend that this approach is limited in two ways. The first criticism focuses on the exclusion of an item discrimination parameter. However, the empirical KCT-R calibrations utilized in this analysis were derived using the dichotomous Rasch model, and item fit statistics were quite acceptable. Guessing a tapping pattern is not a rational possibility on this instrument. The first criticism thus seems moot.

The second concern is the use of R^2 and the way the wide range of difficulties influences its estimation. We know that range restriction [and elongation] influences correlation. The matter under investigation, however, is not a calculation from simulated or restricted data samples, but is a cross-validation of data collected from separate samples spaced over almost 100 years. The ensemble results do not seem to capitalize upon chance, nor are they solely a capricious function of range. As Brody indicates, the results generalize beyond the separate sample means.

It is relevant here to recall Guilford's (1954) suggestion that there are two types of equations. An *empirical equation* is "good for descriptive purposes, but has no theoretical implications" while a *rational equation* is "…developed from known or assumed facts … for how well the data fit the function" (pp. 54–55). This is the same distinction that Thurstone (1959, pp. 5–6, 279–280) made between correlations (as, for instance, of circle diameters taken as indexes of area) as arbitrary parameters, on the one hand, and measures, on the other, which are meaningful in terms of comparing amounts of the substantive phenomenon (area) directly modeled and represented by an equation.

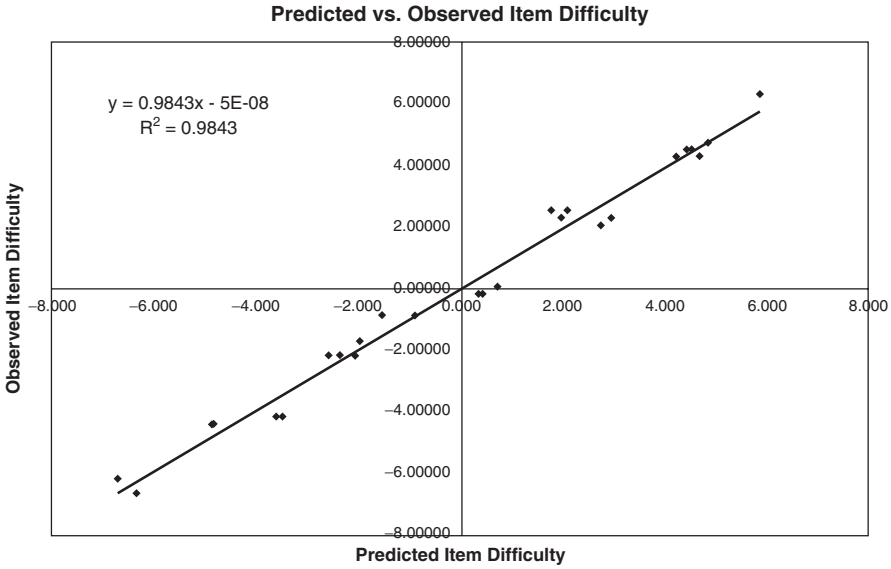


Fig. 6.2 Predicted vs. observed item difficulty

Criticism is justified and holds when all we do with the values computed is use them to describe what has occurred. If an a priori and theoretically sound process is confirmed over multiple time periods and separate samples we have an entirely different matter. Confirmation moves beyond description as theory trumps data. The plot of ensembles confirms the construct theory (simple as it is) for the KCT-R items by virtue of the cross-validation supporting a theory of item difficulty.

Furthermore, we can test the efficacy of the original KCT construct theory as expressed in the equation by generating new items, which are found to take scale positions exactly where theory predicts them to be located (Stenner & Stone, 2003). For an existing instrument, we can generate new items that, when administered to a sample from the same population as previous samples, should calibrate in MITs near where other similarly composed items are already located. This predictability can be useful for building alternate forms, or for generating items "on the fly" for single use applications (Stenner, Fisher, Stone, & Burdick, 2013). For the KCT-R, simply reversing the item taps for the 26 existing items produced an intra-class correlation of 0.97 between the alternate tapping form and the original.

The full range of an instrument might need to be "filled in" where gaps between items exist or where the range must be extended by constructing additional items (Wright & Stone, 1979, 2002). For instance, constructing a KCTR item at item difficulty −2.0 we have the equation:

$$2.164 * 4 \text{ taps} - 10.84 = -2.184 \tag{6.1}$$

Inasmuch as taps cannot occur in less than whole numbers we can expect that 4 taps will work, and also be within the error of measurement. For nine taps we have the equation

$$2.164 * 9 \text{ taps} - 10.84 = 8.64 \tag{6.2}$$

for item difficulty in constructing an item beyond #26, the hardest item of KCTR. Reverses and distance can be added parameters to these equations for greater specificity.

The simplicity of KCT-R is useful for two reasons. First, it offers a simple, easy to understand example that grounds discussion about item construction and generation in a concrete, substantive illustration. Lord Kelvin is quoted by Bridgman (1927) as saying,

> I never satisfy myself until I can make a mechanical model of a thing. If I can make a mechanical model of it I can understand it. (p. 150)

Carnap (1966) further explains this matter,

> The mind works intuitively, and it is often helpful for a scientist to think with the aid of visual pictures. (p. 176)

Kelvin's and Carnap's observations illustrate how important it is to have a simple, visual, and manipulative model to guide one's thinking, even though most aspects of reality are undoubtedly far more complex than any scientific model. In fact, the ideal is to model the process mathematically, and use the equation as a theory *par excellence*.

Stenner and Stone (2003) argue that an algorithm that makes it possible to write an item calibrating at any desired level is clear evidence for construct validity as it demonstrates experimental control over the cause of item difficulty. Factor analysis, in contrast, is generally descriptive. Jöreskog in his seminar at the U of C differentiated his three software programs at that time as exploratory, confirmatory, and LISREL. Ben understood from his work using factor analysis that description was not prediction. He constantly reiterated the need to predict and validate.

6.4 Implications for Automatic Item Generation

The KCT-R illustrates that when the rules for item generation are clear and simple the process is relatively easy. Naturally, most examples are far more complex, and the item theory and variable constructs are much harder to specify and operationalize. But the fabrication process with a complex variable is similar to the process used for a simple one. The complexity of a construct, although sometimes overwhelming, need not be completely frustrating. In science, we must simplify if we are to gain understanding. The variable begins as a line, an arrow, in the direction of more person ability and/or item difficulty. It is formulated from an understanding of the *person*

characteristics thought to express "more" of the variable. Items are then constructed to "test" these item difficulty and person characteristics (Wright & Stone, 2002). The idea that binds these items and persons together stems from what Chomsky (1968) designated the "root" and "deep structure" that underlies language and produced his theory of transformational grammar. This is an "ur-variable" in the sense of foundational and primitive. It is latent.

Bormuth (1970) continued this linguistic pursuit outlining the development of achievement items. Bejar (1990) moved from there to automatic item generation applied to a greater variety of tasks and items, and Bejar (2002) provides a summary of the entire process. The seminal and practical chapters comprising Irvine and Kyllonen's (2002) *Item Generation for Test Development* provide a comprehensive summary of the thinking behind item generation for developing tests. But the major issue pertaining to all complex tasks and variables is to deduce the simple, essential elements which provide the maximum information about controlling item difficulty. Extolling complexity for its own reward is of no value in producing an algorithm, no matter how complicated, for item generation.

Deciding upon what to focus on and what to ignore is essential. Isherwood (1939), a poet, writes, "I am a camera with its shutter wide open, recording, not thinking." Though many philosophers and scientists have sought to emulate Isherwood's approach, consensus has settled more in the direction of accepting that attention is focused or attracted by what is meaningful, though there are diverse perspectives within this broad position (to name just two: Wittgenstein, 1958; Whorf, 1956). For their part, measurement theorists, too, have long recognized "…the odd fact that the language itself which we use in our quantitative description of the world, conditions in a subtle way the image that we obtain" (Falmagne & Narens, 1983, p. 287). And electrophysiological research has lately shown that the perception of color does in fact appear to be affected by language at the physical level (Athanasopoulos, Wiggett, Dering, Kuipers, & Thierry, 2009). William (1956) expressed this matter as to the necessary influence of historical, cultural, and linguistic interests, saying,

> Theories may be very simple, while the phenomena they model do not appear simple … theory covers only one of the interesting factors. (p. 7)

Although item generation may lead to test development, it is essential that this process be closely monitored. Constants and mathematical equations need reviewing, updating, and modification. Deming writes in his introduction to the reprint of Shewhart's *Statistical Method* (1939, 1986),

> There is no true value of anything. There is, instead, a figure that is produced by application of a master or ideal method of counting or measurement. This figure may be accepted as a standard until the method of measurement is supplanted by experts in the subject matter with some other method and some other figure. (p. ii)

Rasch (1964, pp. 24, 2, 3; 1980, pp. 37–38) similarly remarked that models are not meant to be true, but useful. No strategy, method or equation may last forever, and the ever-present task of the scientist is to constantly address and validate those values critical to the process under investigation.

The matter of quality assessment and improvement is vitally important in psychometrics. It must be the passion of psychometrics. Witness the continuing attention of physicists to the determination and/or the substantiation of constants and correction of their values. Youden (1972) writes,

> Much ingenuity and labor is expended upon the experimental determination of the values for physical constants. Whenever independent studies have been made a critical evaluation is in order. The data evaluation has two objectives: first to pick a 'best' value for the constant, and second to set some limit to the error in this best value. (p. 1)

An interesting table and figure from Youden (1972), reproduced in Stigler (1999, p. 365) illustrates the problem for the physical sciences in determining the astronomical unit and measuring the speed of light. Psychometrics must address similar problems. If the physical unit is frustrating to science, imagine the work ahead for establishing units in the social sciences (Stenner et al., 2013).

Similar to the problem of error in constants, item generation techniques require scrutiny to determine if the conditions hold or circumstances change. For the KCT-R, differences in gender have not been demonstrated (Stone, 2002a). Richardson (2003), who reviewed the literature on Knox's Cube Test, indicated likewise. He also cites an interesting study by Martin, Franzen, and Raymond (1996), who used Stone and Wright's (1980) version of the Knox's Cube Test. For 23 patients who suffered left hemispheric cerebral-vascular accidents and 40 patients with right hemispheric lesions, there was no significant difference in their performance on the KCT, though the digit span subtest of the WAIS-R produced large and statistically significant differences. It appears that hemispheric damage may not disrupt performance on the KCT although the digit span subtest indicated a diagnostic difference. The KCT appears to measure an aspect of cognitive functioning that is somewhat different from what digit span measures. It is not simply visual vs. auditory in these two measures.

Recent KCT-R samples from Japan and India (Stone, 2002a) have shown no overall differences when compared to American samples. These studies suggest the differential effects of person ability, namely attention and immediate memory, remain the prominent person characteristics. The median performance on the KCT-R occurs at age 18, but the 84th centile was reached at age 13 by some subjects.

The quality of an item can be influenced by its stability or its transience. Stability can be upset by either or both the internal and external aspects of an item or its construct; those item characteristics that support or frustrate item invariance. If the item difficulty equation for the KCT-R is not stable due to any problems encountered from the sequence of taps, reverses or distance, the entire process is upset and suspect. This would reflect a quality issue internal to the instrument itself.

With some items, however, the critical factor may be external; improper test administration, local environmental effects, etc. may occur. Otherwise useful items may become adversely affected by external situations. For instance, a once useful item from the now out-of-date version of the WISC-R, from 1974, asks, "Who was Charles Darwin?" Might the difficulty level of this item have changed as a result of the continuing evolution/creationist debate? Were the item asked

today compared to thirty years past would there be a noticeable difference in difficulty? What may have been a more difficult item about who was Darwin thirty years ago may not be at the same level of item difficulty today because of media attention to the debate about creation vs. evolution. The results of item generation methodology can become a casualty to external events. Internal and external matters can destabilize the item generation process.

There are psychometric strategies for dealing with these item generation problems, but the process of item generation can operate no higher than the level of construct stability that has been achieved and maintained. We can never be absolutely assured of exactly what makes any item difficult through the thick and thin of reality. Internal and external matters can produce changes. We can only be alert to the possibility of change occurring. Staying alert requires a process of continuous quality control, and ever present attention to whatever internal and external sources might upset the process.

Concern for the person to be measured supersedes the item in order of attention. Item generation is not a scene from *Das Glasperlenspiel* "The Glass Bead Game" (Hesse, 1969). We do not generate items solely for intrinsic pleasure. Item generation is not a psychometric exercise that has no purpose save its own perpetuation. We should never lose our focus on the person. We should first determine and resolve what person attributes are of maximum importance, and which ones can be eliminated from our concern. These attributes become the focus of attention, and then we address the issues of item generation, not the other way around.

The conundrum is that persons and items constitute a two-way frame-of reference required for our deliberations, procedures, and calculations. One dimension from the item-person frame cannot be fully resolved alone. Each one must be derived in the context of the other. This requires item generation to constantly address its purpose, and validate its process against success in measuring persons. Shewhart's model (1986; Stone, 2002b) for quality assessment requires

$$\text{Specification} \rightarrow \text{Production} \rightarrow \text{Inspection}$$

Item generation is no less concerned with this production process than any other fabrication activity. Messick (1992) argued that a c*onstruct-centered* approach,

> would begin by asking what complex of knowledge, skills or other attributes should be assessed … what behaviors or performances should reveal those constructs … what tasks or situations or situations should elicit those behaviors? (p. 17)

The task of item generation is to respond to these questions. One hundred years ago Knox addressed these matters to solve a difficult problem that had to be faced. His instrument satisfied that need and today serves as an inspiring guide to construct validity via item generation.

6.5 Concluding Comments

In *Best Test Design*, the KCT served as a concrete learning example even as it continues to serve in clinical work, and as one among many other analytic examples given in the WINSTEPS Manual (Linacre, 2014). It has been personally interesting to have studied individuals and data from using Knox's Cube Test in my career with Ben Wright, and to have estimated parameters first by hand, and subsequently by a succession of software programs. The next step is to present the KCT items in a continuous format using computer administration.

References

American Psychological Association. (1954). Technical recommendation for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement, 51*(2, Part 2), 1–38.

Athanasopoulos, P., Wiggett, A., Dering, B., Kuipers, J.-R., & Thierry, G. (2009). The Whorfian mind: Electrophysiological evidence that language shapes perception. *Communicative & Integrative Biology, 2*(4), 332–334.

Arthur, G. (1947). *A point scale of performance tests*. New York: Psychological Corporation.

Babcock, H. (1965). *The Babcock test of mental deficiency*. Beverly Hills, CA: Western Psychological Services.

Bejar, I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14*, 237–245.

Bejar, I. (2002). Generative testing: From conception to implementation. In H. Irvine & P. Kyllonen (Eds.), *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Bormuth, J. (1970). *On the theory of achievement test items*. Chicago, IL: The University of Chicago Press.

Bridgman, P. (1927). *The logic of modern physics*. New York: Macmillan.

Brody, T. (1993). *The theory behind physics*. New York: Springer.

Carnap, R. (1966). *An introduction to the philosophy of science*. New York: Dover.

Chomsky, N. (1968). *Language and mind*. New York: Harcourt Brace & World.

Cronbach, L., & Meehl, P. (1955). *Psychological Bulletin, 52*(4), 281–302.

Davis, A., Gardner, B., & Gardner, M. (1941). *Deep south*. Chicago, IL: The University of Chicago Press.

Falmagne, J.-C., & Narens, L. (1983). Scales and meaningfulness of quantitative laws. *Synthese, 55*, 287–325.

Guilford, J. (1954). *Psychometric methods*. New York: McGraw-Hill.

Henry, W. (1956). *The analysis of fantasy*. New York: Krieger.

Hesse, H. (1969). *The glass bead game*. New York: Holt, Rinehart & Winston.

Irvine, H., & Kyllonen, P. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Isherwood, C. (1939). *Goodbye to Berlin*. New York: Vintage.

Linacre, J. (2014). *A user's guide to WINSTEPS*. Chicago: MESA.

Martin, R., Franzen, M., & Raymond, M. (1996). Effects of unilateral vascular lesions and gender on visual spatial and auditory verbal attention. *Applied Neuropsychology, 3*, 116–121.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Pintner, R. (1915). *A scale of performance tests*. New York: D. Appleton.

Rainwater, L., Colman, R., & Handel, G. (1962). *Workingman's wife*. New York: MacFadden.

Rasch, G. (1964). An individual-centered approach to item analysis with two categories of answers. Unpublished ms.
Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press. (Original publication 1960).
Richardson, J. (2003). *Knox's cube imitation test: A historical review and an experimental analysis*. Unpublished manuscript. The Open University, Walton Hall, Milton Keynes MK76AA, UK.
Shewhart, W. (1986). *Statistical method from the viewpoint of quality control*. New York: Dover. (Original work published in 1939).
Stenner, A. J., Fisher, W. P., Jr., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology, 4*(536), 1–14.
Stenner, A. J., & Smith, M. (1982). Testing construct theories. *Perceptual and Motor Skills, 55*, 415–426.
Stenner, A. J., & Stone, M. (2003). Item specification vs. item banking. *Rasch Measurement Transactions, 17*(3), 929–930.
Stigler, S. (1999). *Statistics on the table*. Cambridge, MA: Harvard University Press.
Stone, M., & Wright, B. (1980). *Knox's cube test*. Wood Dale, IL: Stoelting.
Stone, M. (2002a). *Knox's cube test—revised*. Wood Dale, IL: Stoelting.
Stone, M. (2002b). Quality control in testing. *Popular Measurement, 4*(1), 15–23.
Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: The University of Chicago Press.
Thurstone, L. L. (1959). *The measurement of values*. Chicago: University of Chicago Press, Midway Reprint Series.
Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. In: J. B. Carroll (Ed.), (Foreword by Stuart Chase). Cambridge, MA, New York, and London: Published jointly by The Technology Press at MIT; John Wiley & Sons, Inc.; and Chapman & Hall, Ltd.
William, J. (1956). *The compleat strategist*. New York: McGraw-Hill.
Wittgenstein, L. (1958). *Philosophical investigations (G. E. M. Anscombe, Trans.)* (3rd ed.). New York: Macmillan.
Wright, B. D., & Stone, M. (1979). *Best test design*. Chicago: MESA.
Wright, B. D., & Stone, M. (2002). *Making measures*. Chicago: Phaneron.
Youden, W. (1972). Enduring values. *Technometrics, 14*(1), 1–11.

Chapter 7
The Influence of Some Family and Friends on Ben Wright

John M. Linacre

Abstract The author's close association with Ben Wright from 1983 to 2001 enables the author to recount Ben's anecdotes about his relationships with his family, colleagues. The early demise of his father guided Ben's somewhat awkward relationships with authority figures. His experiences working for future Nobel-prize winners in experimental physics defined his view of what is good measurement. His childhood educational experiences provided examples of what he wanted, exemplified by the Little Red School House, and did not want, exemplified by The Hill House, in his own teaching methods. Other influences mentioned include Frank Chase, Leonard Jimmie Savage, Georg Rasch and Charles Sanders Peirce.

7.1 Getting Acquainted

Ben Wright and I first met late in 1983 or thereabouts. For some reason, Ben took an immediate liking to me—I reminded him of Bruce Choppin, Ben's first "Rasch" student, who had recently met an untimely death. Though we never met, Bruce and I shared much in common: English origin, educated at Cambridge University, along with some math and computer background. But Bruce had a "reckless courage" (Wright, 1985) that placed him above ordinary mortals.

In the early 1980s, a Federally-funded project, for which I was data processing manager, was running into difficulties. Thousands of tests had been administered to children, but the data were messy—observations were missing. There were dichotomies and rating scales. Our project director visited the top test and measurement experts. Only Ben Wright knew how to make sense of the mess. So Ben wrote out the algebra and I implemented a Rasch estimation procedure that would run on a personal computer with a hard disk (then a recent innovation). It was robust against

J.M. Linacre (✉)
Winsteps.com, 9450 SW Gemini Dr # 27615, Beaverton 97008, OR, USA
e-mail: mike@winsteps.com

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_7

missing data and handled various response formats. We integrated it with a spreadsheet to ease data entry and to produce graphical displays.

When Ben would visit his retired mother in her apartment in Greenwich Village, New York, I would come down from my home in Connecticut. We would spend happy hours together discussing enhancements to the computer software. Soon he was persuading me to enter a Ph.D. program at the University of Chicago, and in 1986, I did.

Ben indicated that my studies would be completed in a year. A friend, a Dean at another University, warned me to expect at least 3 or 4 years. In fact, thanks to encouragement and financial support from the Spencer Foundation, I was done in 3 years. But that was just the start. I became Ben's perpetual student, sitting in all his measurement and questionnaire design classes from then until that sad day, October 30, 2001 when Ben was struck down by a cerebral incident.

7.2 Some Background on Ben

During those years, Ben related many autobiographical anecdotes to his classes and colleagues. He also put various incidents into writing. Here is a compilation of some of those. Treat them as impressions rather than strict facts, because Ben himself was somewhat inconsistent in his account of his own life.

Benjamin Drake Wright was born in Wilkes-Barre, Pennsylvania, March 30, 1926, the eldest child of Dorothy Lynde Wadhams Wright, a New York socialite, and New York banker Harold St. Clair Wright. At first it seemed that his path through life would be strewn with roses, but this was not to be. In an autobiographical section of "Kinesthetic Ventures" (Bouchard & Wright, 1997), Ben recalls how, simultaneously in his early life, he lost his favorite toy (an orange engine) and was "abandoned" to a nurse, who substituted for his mother as his care-giver after the birth of his brother, Raymond.

In later years, it seemed that Ben diligently, sometimes ferociously, safeguarded his ideas. Indeed the only sure way to persuade Ben to relinquish his own position on a matter, his current "favorite toy," is to demonstrate to him a new and better toy. Before long Ben will be playing with the new toy. If he discovers that the new toy is indeed better than his own, he adopts the new toy as his own—the old toy forgotten. This behavior can surprise Ben's antagonists. They suddenly find Ben is a stronger advocate for their position than they are themselves. On occasion this has antagonized Ben's antagonists yet further. In his search for scientific meaning, Ben ignores an implicit maxim of Academia, "Do be my enemy—for friendship's sake." (Blake, *To Hayley*).

His early life was a contradiction. He loved the primary school he attended, the "Little Red Schoolhouse" in Greenwich Village, New York. To Ben it was idyllic. As he often related in class, the children chose what to study. The staff facilitated; they did not discipline. This school became the model for what Ben wanted his own classes to be. But Ben's relationship with his father was negative. His father was always criticizing him. Ben could never do well enough to satisfy his father, no matter how hard he tried.

Then, in 1936, when Ben was age 10, his father died unexpectedly. This left an unresolved conflict in Ben's life, physically manifesting itself as a propensity to stoop when he walks. Ben's family situation and life changed. Ben himself was never able to remedy his relationship with his father. Sometimes it seemed that this disjunction overflowed into his relationships with other authority figures in Ben's life.

His mother was very progressive, so Ben was psychoanalyzed—the introduction to his lifelong devotion to Sigmund Freud. Later Ben became a Freudian psychoanalyst. One of the requirements for this was to have been psychoanalyzed. In fact, Ben was psychoanalyzed twice during his life. But, true to his own approach in many areas, Ben made his own improvements to Freud's theories.

In 1939, Ben's mother left New York and taught herself to manage a thriving dairy/chicken farm near Stroudsburg, Pennsylvania. So, in 1939, Ben entered The Hill School in Pottstown, Pennsylvania. This prestigious school Ben perceived to be excessively disciplinarian. He hated it. This became the antithesis of what Ben wanted in his classes.

One summer, probably in 1942, Ben was with some relatives on a beach in New Jersey when they encountered Albert Einstein. It seemed someone in the party knew Einstein so they stopped and chatted. Ben doesn't recall what the conversation was about, but did notice the Einstein was a kind person. Ben feels that this meeting had no impact on his life, but maybe ….

In 1944, Ben graduated from High School having already enlisted for naval officer training. He was sent to Cornell University. He perceived himself to be lucky because some of his classmates, who went into the ranks, were given their basic training and sent to the war in the Pacific. They were dying while Ben was studying. At Cornell, Ben considered studying electrical or mechanical science, but such courses had so many officer cadets enrolled that they were marched to and from lectures. Ben detested that, so he signed up for physics, an unpopular course. (Did Ben's meeting with Einstein influence that choice?)

In 1946, Ben's mother entered Teacher's College, Columbia, to earn an M.A. She ultimately became a Professor of Education and Psychology at New York University before retiring in 1969. She died in 1995 at age 93.

7.3 On to Chicago

In 1947, the war was over and Ben went to work at Bell Laboratories in New Jersey, working for future (1964) Nobel Laureate Charles H. Townes on microwave absorption spectra. In 1948 Townes accepted a position at Columbia. He invited Ben to join him there, but Ben was already committed to the University of Chicago–the exciting place in physics, particularly since Enrico Fermi's (1942) chain reaction in a squash court under the west stand of the disused football field.

Arriving in Chicago, Ben worked for Robert S. Mulliken, another future (1966) Nobel Laureate, on ultra-violet absorption spectra. The work was painstaking. The same experiment was done many times until it was done correctly. Ben quite enjoyed

it. It taught him the importance of good measurement, and that nearly all data collected in scientific experiments is rejected as hopelessly flawed. But Ben also discovered that the people in that field were really colorless. He decided that, if he didn't want to become like one of them, he had better change his life course.

His first look was at the Committee for Social Thought, but they were never in their offices. Then in 1948 he encountered the Committee on Human Development and Bruno Bettelheim. He also married Claire Marie Engelman, a nursery school teacher. In 1950, Ben became a psychotherapist at the Orthogenic School. According to Ben, Bettelheim was treating children who had been written off by others as untreatable. Bettelheim was working in uncharted territory so, of course, mistakes were made. It was physically exhausting and emotionally taxing work.

Soon it was 1957, and Ben was beginning to feel burned out with dysfunctional children. Serendipity struck. According to Ben, Egon Guba, later a distinguished figure in the world of qualitative research, was teaching the introductory statistics class in the Education Department. He wanted a pay rise, so he tried to pressure Frank Chase, the Department Head, by saying that he had a received an offer from Ohio State University. Frank Chase said "Congratulations on your new position. Goodbye!".

At short notice, the Education Department needed a statistics instructor. Bruno Bettelheim knew of Ben's background in mathematics and physics, and so recommended him. Ben got the job. Ben's previous exposure to statistics was only cursory, comprising a statistics course by William Stephenson (of Q-sort fame) and a course on probability theory by William Feller. But Ben was good at math. So he read through the statistics textbook. Soon he discovered various mistakes in it, and, in class criticized the book and made fun of it. He proceeded to teach his own version of what statistics should be.

Ben later discovered that Egon Guba didn't like that textbook either. But the other Education faculty were furious at Ben's approach to statistics and wanted Ben fired. Frank Chase called in Leonard "Jimmie" Savage, head of the Statistics Department and a leading proponent of Bayesian methodology, to adjudicate. Savage came down on Ben's side, and Ben kept his job. In retrospect, Ben said his reaction to the textbook was driven by his own insecurity in this new area, and his need to assert himself. This reaction was observed, on occasion, throughout Ben's life, affecting his relationships with Georg Rasch and Fred Lord.

Ben's contact with Savage also led to his meeting the "father of modern statistics", Ronald A. Fisher. Fisher had been "put on the shelf" in Britain. During WWII, many of Fisher's students contributed to the British war effort in data analysis, code breaking and military logistics, but Fisher himself was sidelined. This continued after the war, so Fisher accepted a position in Adelaide, Australia. On his way there, he went through Chicago. Savage, Ben, and the statistics faculty had lunch with him at the Faculty Club, and then went downtown to hear him lecture. Fisher was critical of the way that the 0.05 significance level had become a gold standard for hypothesis tests. He stated that the significance level should be set according to the nature of the hypothesis, and in any case, it was not one occurrence, but a systematic observance of that level which motivated acceptance/rejection. This accorded with Ben's own view of statistical tests. Fisher proceeded to Adelaide, but died soon after arriving there.

In his early years as a statistician Ben needed money to support his growing family, so he supplemented his meager salary with research work. Ben's salary was always at the bottom of the scale and his promotions were few, because of his habit of alienating the University hierarchy, which was perhaps also a reflection of his relationship with his father. In fact, in 1962, the Education Faculty opposed granting Ben tenure, so Frank Chase, Ben's academic hero, arranged a Faculty meeting in the middle of summer, when few faculty were present, and railroaded through the granting of tenure to Ben. Over the years, Ben often remarked that he didn't want to receive honors. He would recount the story of one distinguished Professor of his acquaintance who was dragged out of retirement to receive some University honor. The stress of it all killed that Professor soon afterwards. But, when honors did come uninvited, Ben certainly relished them.

Ben's paid research work involved factor analysis of market research data. Earlier William Stephenson had taught Ben how to do factor analysis by hand. Ben also got to know Leon Thurstone, the leading factor analyst, who did his work in a big basement room in Green Hall. His large table there was covered with sheets of paper. But, not long after Ben got to know him, Thurstone moved to North Carolina. Ben himself began to use the University's new IBM 7090 computer in the basement of the Hall of Administration. During the day, the computer did the University accounts, at night it was used by the academic researchers. Ben wrote his own factor analysis program, which was widely used by other faculty. In fact, Ben and his computer programs were among the most frequent computer users.

Ben's other interests in psychology were leading him into teacher training and editorship of *The School Review*. He was proud of their colorful covers (which tied in with the factor analytic research into color he was doing for a tobacco company).

Ben and Claire were living on South Drexel in a now-demolished apartment block. Jimmie Savage was their neighbor, so they became quite familiar. Savage encountered Georg Rasch, and told Ben that he would invite Rasch to Chicago if Ben was interested. Ben said he was, so Rasch arrived. Initially, Rasch was a distraction. Ben was more interested in factor analysis than item analysis. But Rasch had written papers on factor analysis, so Ben discussed the instability of factor structures with him, a problem of major concern to Ben's clients. This motivated Ben's interest in the other aspects of Rasch's work. Ben was the only participant to sit through Rasch's entire course of lectures—more out of sympathy for Rasch than anything else, perhaps.

7.4 Psychometric Innovations

Once introduced into the item analysis world, Ben encountered Fred Lord and Louis Guttman. Louis Guttman had made his name during WWII with his scalogram analysis–a deterministic theory. Louis and Ben had meals together a number of times, and Ben tried to convince Louis of the value of a stochastic version of the scalogram, i.e., a Rasch model, but Louis could not be persuaded.

Ben seemed to have better success with Fred Lord. Initially Fred Lord was not a strong advocate of any particular IRT model, he just wanted something that would work. Ben discussed his own research with Fred, and Fred was interested. Fred requested a copy of Ben's Rasch computer program, and Fred's computer guru tried to get it to work. But it wouldn't work properly. Then Fred's empiricist approach started to conflict with Ben's idealism.

By now, Ben himself was experimenting with new ideas. In 1967, Ben's students included Bruce Choppin and Nargis Panchapakesan, a physics Ph.D. from India. Nargis was just looking for something to do while her husband did his studies. Someone directed her to Ben because of his physics background. Ben persuaded her to do another Ph.D. in Education. The three of them went to work on item analysis methodology.

One outcome of these efforts was a program implementing a model with two item parameters (2-PL), bringing in discrimination alongside difficulty, in the manner of what is now known as Item Response Theory. Ben liked it, but they couldn't get it to work. Finally they gave up on it. Later, with Graham Douglas, Ben proved that a 2-PL computer program couldn't work without the introduction of arbitrary constraints. This confirmed Ben's skepticism about Fred Lord's approach.

Another outcome was unconditional (UCON, JMLE) estimation. This somewhat alienated Ben from Georg Rasch. Rasch advocated "statistically consistent" but computationally awkward and limited methods. In this case, Ben was the realist advocating simple, practical methods. Throughout his career, Ben has faced similar confrontations. On the one side Ben argues with the data-driven empiricists. On the other side he argues with the theory-driven perfectionists. Ben's philosophy accords with Einstein's statement that "Everything should be made as simple as possible, but not simpler."

A third outcome was a mixture of factor analysis and a method taught by Georg Rasch. Rasch had constructed, out of his unidimensional analysis of dichotomous data, a multi-dimensional technique for analyzing polytomous data. But Ben didn't like it. His physics background mandated "measure one thing at a time." So Ben devised a method in which Rasch's multidimensional rating scale measures were factor analyzed, and their first factor became Ben's desired unidimension. It was a sloppy approximation, but "good enough for government work," which was always Ben's response when perfectionists pointed out flaws in his work. A few years later, the research of Erling Andersen and David Andrich rendered this method of analyzing polytomies obsolete.

7.5 Later Years

We have now reached 1980, and Ben himself has become the father-figure influencing generations of young students. Each student has imparted a little to Ben, but Ben has imparted far more to them. He was particularly impressed by Isabel, a bored suburban housewife. She had nothing to do but drink coffee with her equally bored

friends and gossip. So she signed up for a Masters degree in Social Science at the University of Chicago. A requirement was a statistics course. Ben's course was perceived to be the least arithmetical, so many qualitatively-oriented students signed up. Isabel was attracted by Ben's psychoanalytic approach to interpreting item difficulty hierarchies and so applied Rasch analysis to the participants in her coffee klatch. She measured the degree of neurosis of each participant and the diagnostic strength of the conversational indicators. Ben perceived this to have been the most creative work done by one of his students in his classes—it reinforced his perception that an introductory course did not have to consist of canned repetitions of trivial examples.

The major influence to enter Ben's life in later years was Charles Sanders Peirce. Peirce died in 1914, before Ben was born, but Ben would dearly love to have met him, and regarded him as a kindred spirit and friend. Peirce was a physicist, and a mixture of theoretician and practitioner, philosopher and scientist. Peirce also was a genius who was not properly recognized. Peirce lost his position at Harvard over an indiscretion which today would not even be remarked upon. Ben felt that Peirce's ideas were appropriated by Dewey and James without due credit. So Ben felt that Peirce was an unappreciated victim of the system, and, on occasion, he saw parallels in his own career. It was a joyful day when Ben realized that Peirce himself had formulated a Rasch-type log-odds model in 1878 (Linacre, 2000). All along, Ben was convinced that Rasch measurement was that vein of gold for which philosophers across time, at least since Plato, have sought diligently. Suddenly his latest hero, Peirce, was found to have discovered it! The validity of Ben's life's work was compellingly confirmed for him in this.

References

Bouchard, E., & Wright, B. D. (1997). In M. Protzel (Ed.), *Kinesthetic ventures*. Chicago: MESA Press.

Linacre, J. M. (2000). Almost the Peirce model? *Rasch Measurement Transactions, 14*(3), 756–757. Retrieved from http://www.rasch.org/rmt/rmt143b.htm.

Peirce, C. S. (1878) *Illustration of the Logic of Science by C.S. Peirce*, Assistant in the United States Coast Survey. Fourth Paper: The Probability of Induction. Popular Science Monthly, pp. 705–718.

Wright, B. D. (1985). Memories of Bruce Choppin. *Evaluation in Education, 9*(1), 7–8.

Chapter 8
Things I Learned from Ben

Mark Wilson

Abstract In this chapter I briefly describe four things I learned from Ben Wright.

8.1 Introduction

After I had finished my dissertation under the leadership of Ben Wright at the University of Chicago (in 1984), I gave Ben a copy of the dissertation manuscript that I had had bound in the rather sombre maroon color favored by the University. I inscribed in it my thanks for his leadership and kindness to me. I noted in that inscription that Ben had been the "best reader I had ever known." I think he was a little disappointed at this, which might have seemed minor praise. But, in fact, wound up in that small phrase was a tribute to his exceptionality, and to his generosity of spirit. It is late in coming, but, in these next pages, I will attempt to explain what I meant by those few words. Here are, just a selection of, the things I learned from Ben Wright.

8.2 Find People Who Have Good Ideas, Listen to Them, and Work with Them (If You Can)

When Ben was leading courses (it's hard to say he was merely lecturing, as his courses were always structured like a serial novel, and each individual talk had its own pace and drama), one of his tactics was to use quotes from people who he thought were interesting. He would use these quotes to focus the class's attention on a provocative idea, then he would engage in a bantering (sometimes even heckling)

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-3-319-67304-2_16

M. Wilson (✉)
Graduate School of Education, University of California, Berkeley, Berkeley, CA, USA
e-mail: markw@berkeley.edu

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_8

conversation with the person quoted. He would take the position drawn in the quote, and run it up against another, extending the comparison to illustrate the tensions between the positions. Sometimes it seemed a bit far-fetched—"Did they really mean all of that?" one would wonder. Well, maybe not. Did Ben really espouse those (far-fetched) ideas? Probably. But what had happened was that the class-members were not just listeners any more, they were imagining themselves in that conversation, each was speculating about their own position in that conversation. Not many of us had the gumption to voice those positions we had found—it was pretty overwhelming being exposed to such strong ideas. Certainly they were not like the careful, safe and circumscribed lectures I had experienced in my previous university courses.

In my own research career I had an experience where listening to experts was a key to dealing with the situation. I was on a NRC Committee examining the role of recent developments in cognitive development on assessment (see NRC, 2001, for the final report). The Committee was composed of a large group of experts, with just a few (token?) measurement experts involved. The measurement group was asked early on to describe what we thought could be our contribution to the work of the Committee, and as a part of that, we had each Committee member say what they thought was the source of the problem that the Committee was charged with addressing. The majority said that they thought the problem lay with the limitations imposed on assessment by measurement experts and the testing industry. This led the measurement sub-group to speculate that our true role was possibly to be "sacrificial lambs," absorbing the criticisms of the rest of committee, and helping them understand how better to aim those blows.

What we needed to do was to explain the range and possibilities of measurement, and how the techniques of measurement could be applied in ways that did not impose limitations, but in fact, informed new developments instead. In order to get to a situation in the Committee where we could turn this around, we needed to *listen* to those critics—each a genuinely brilliant expert in their own domain of scholarship—we needed to listen not just to their general criticisms, but also to understand what was the special perspective from their disciplines, and figure out how to express our insights about the potential of measurement to be an effective tool for testing, and not the limitation that they saw. I'm not sure that we really managed to do that for all of those Committee members, but we did succeed for some.[1] In fact two of those Committee members became collaborators in my own research, and continue to actively collaborate with me today, over 10 years later. It seems that some of them were good at listening too!

8.3 Respect Your Own Ideas

In contrast to the first "lesson learned" above, this second one concerns one's view of oneself—do not under-appreciate your own thoughts, your own creative ideas. We can each be our own worst enemy, being the very first to critique our own

[1] You can read the results of our efforts yourself in Chapter 4 of the NRC report (NRC, 2001).

creations. It is important to let your ideas have free rein for a time, so that they aren't smothered by worries about how sensible or true they are.

Ben was truly astounding sometimes with the things he would say. For example, one of the first things I heard him say in a class was "a survey is a conversation." This seems pretty silly at first sight—how can a survey be a conversation when one side is a piece of paper, and the other is a person checking off alternatives? Doesn't a survey lack that essential feature of any conversation—the interactivity between the two parties in the conversation? Surely Ben must have thought that to himself when he first thought it—after all, his conversations have always been so vibrantly interactive.

And, then it became clear, as he followed up that initial remark, that there is indeed a sense, an important sense, in which a survey constructor must seek to engage the survey respondent in a virtual conversation. He turned something that at first seemed implausible, into an important argument, and one that had an attention grabbing "hook" to start with. (See the materials in Wright, Enos, Enos and Linacre (2001) for an argument for this idea.)

In developing my ideas for a dissertation, I was interested in finding substantive theoretical work on cognitive structures that I could wrestle with from a measurement point of view. I was hard-pressed to find the right combination of amenable cognitive theory with available associated data. Eventually, I found that the topic of Piagetian-like stages had been around long enough for researchers to have collected some reasonably large data sets (even though Piaget himself never did gather data in this form). And, of course, who could doubt the importance of Piaget's "stages"? (In fact, at about the same time, developmental psychology was going through a pretty whole-hearted rejection of much of Piaget's work!) But devising the right statistical structure was difficult, and I ended up completing the dissertation (Wilson, 1984) without really being satisfied with the statistical expression I developed for relating the stages to the underlying metric.

But, with Ben's encouragement, I kept at it, and shortly after I had finished, I found an alternative formulation based in mixture models (Mislevy & Verhelst, 1990) that gave me the frame with which to complete my conception of the "saltus" model (Wilson, 1989)—linearly-constrained mixture models. I was sure I had an interesting problem, but I had to persist through some very rational doubts about the estimability of the model to find a better concept.

8.4 Respect Your Own Doubts

The third thing I learned also relates to one's view of oneself. Do not get overwhelmed by somebody else's ideas: Learn to listen to your own doubts about what someone is saying. If something sounds a bit dubious, then pursue that doubt. Ben has been a very thorough doubter of other people's ideas. In fact, I sometimes think that this sometimes led him to doubt other people's integrity—how could they possibly say that: don't they doubt themselves?

His well-known antipathy for the Two-Parameter model (2PL) is a formidable case in point. It seems to so many that it is simply *de rigueur* that more complex models are always "better." But Ben stuck to his doubts, and expended many years and many pages of manuscripts expressing his concerns about that common mistake. Instead, he turns that common misunderstanding on its head: Why would anyone want to create sets of items that were inconsistent in their discrimination (slope or steepness) parameters? Can't they see that it leads to all sorts of problems and difficulties? Why not use the statistical model as a principle of test development? (See Wilson (2003) for more on this.)

In my own research I have found that my doubts are a signal guide for what to do next. This has been particularly prominent in my more applied work. When I have looked carefully at what is used by policy-makers to justify their choices in the area of testing and assessment, it has been regularly the case that I feel doubts about those justifications. One place I have seen this clearly is in the historical approaches to cut-score setting (also called "standard-setting").

My interest in this area was sparked by my consideration of the historically-dominant (though much-modified) method: the Angoff Method (Angoff, 1971). Here, people who know that it is very hard to gauge the difficulty of items based on their contents (i.e., psychometricians), recruit other people,[2] who do not know that this is indeed very difficult, into a standard-setting committee. In their deliberations the committee-members use their intuitions about item difficulty to complete certain tasks. Then the psychometricians study the success (or lack of success) of the committee members, and deem the result to be acceptable or not.

From the very first, this process had seemed dubious to me. Why study people doing a job for which they have no special qualifications? Surely the task is not to study people in this unfortunate position, but to devise ideas and tools that will help them, and lead to more useful decisions? This logic has led to the development of the Construct-Mapping procedure (Wilson & Draney, 2002), which is still not perfect (so I still have my doubts), but indeed provides more information for standard-setters than other procedures (Draney & Wilson, 2009).

8.5 Go for It

When an idea is a good one, then it is worth investing your time and effort into it. At the time I first met Ben, when I started as a graduate student in Chicago, I was expecting to find a standard campus-variety academic: Concerned about learning and evidence, and active in the domains of learning and academe. Yes, he was that. But that was only a small part of his field of action. He added to these common academic areas a charismatic pursuit of an idea that bridged theory and the real world, and that engaged people passionately within both academe and public policy

[2] Typically these are teachers, or other people who have professional interests in testing but are not actual measurement specialists themselves.

domains, including educational testing and medical measurement. He pursued this with relish, smiling most when it seemed most risky: He used not just academic tools, but all the wiles of his intelligence and personality.

Seeing Ben's wonderful energy and commitment has meant a great deal to me. His example has freed me to also step outside my academic roots, and enter into domains that are risky. At first in my academic career, I spent my time writing papers about somewhat obscure statistical matters applied to measurement issues, laying down sufficient journal citations for a secure, tenured position. But, to me this seemed like small chips compared to what Ben was up to. So, I started working with a real-world problem: How do you develop assessments for an educational curriculum? What do you look for in the curriculum as the target(s) of the measurement? How do you communicate results back to teachers and students? It took a decade to get a single paper out of that (Wilson and Sloane 2000)! (Of course, I had to keep my academic hobby topics going too, or they would not have given me any promotions.)

Questions like these have kept me busy for over 20 years now. I have gotten to the point where I think we have some new answers to these questions. I found that I could not afford to wait until the curriculum was completely developed in order to develop assessments for it (as the assessment information is an essential part of the curriculum development). I found that I had to go out beyond my "safe zone" of statistical modeling to respect the intents of the curriculum developers (so I still keep my hand in the statistical modeling game). I found that I could take advantage of one the intellectual tools that Ben developed to help non-specialists to interpret complex assessment results (i.e., the "Wright Map"—see the Appendix 8.1). I cannot yet declare success on this endeavor, but I am persisting (see, for example, Wilson, 2012), and it certainly is interesting!

8.6 Conclusion: What It Means to Be a "Good Reader"

So now you can see that when I said Ben was my "best reader," I meant a lot lot more than perhaps was obvious. I meant that he had startled me, and inspired me, with other people's ideas (and his own). I meant that he had shown me what it meant to take yourself and your ideas seriously, and had encouraged me to do just that. I meant that he had fostered my doubts, and lived out his own. I meant that he had shown that the academic life is definitely worth living, but you probably have to escape the academic world in order to fully enjoy it. Thank you Ben.

Appendix 8.1: The "Wright Map"

I had heard the term "item map" being sometimes used to describe the representation of items and persons on the same graph. I am not sure the origin of that term, nor of the idea. But I knew that, for many years, Ben Wright had championed this

approach to interpreting the results of measurement analyses, and also that he had made significant contributions to that approach, including enhancements and adaptations such as kidmaps, fit maps, maps for polytomous items, etc. It seemed to me that, in fact, Ben had made his most significant contributions to measurement in the area of conceptualizing measures, and interpreting the results of measurement analyses, and that his central tool in doing so were these (many forms of) item maps. In addition, I knew of no one else who had made an equivalent contribution, especially not in terms of item mapping. Hence, I coined the term "Wright Map" in honor of Ben Wright and his very deep contributions to measurement.

This was at about the end of 1999 and the beginning of 2000. After that, I used the term in my class (EDUC 274A, "Measurement in Education and the Social Sciences I" at UC Berkeley) to get used to the sound of it—the students seemed to find it quite a useful term. As I was at that time drafting the text of my book *Constructing Measures* (Wilson, 2005), it became embedded in the text. The first time I used the term in a formal presentation was at the International Conference on Measurement and Multivariate Analysis, Banff, Canada (Wilson & Draney, 2000). I also used it at the first ICOM conference in Chicago (Wilson, 2001): that is the first time Ben Wright heard it, as he was in the audience (he told me he was very moved, and flattered).

Some might be surprised that Ben didn't invent the term himself, as he was thought far from modest in most matters. But I believe he was indeed quite modest in formal matters, and was delighted to hear his name being used in this way. As far as I know, the first time the term appeared in print was in the Proceedings from the Banff conference (Wilson & Draney, 2002). The second presentation was also published (Wilson, 2003), and a version of it is also included in my *Constructing Measures* book.

Generally, I have found that people have welcomed the term—no one has ever objected to it, in my hearing, though, of course, they might not do so directly to me. It seems it has gained some currency: I searched for it in Google[3] just the other day, and got 1180 hits. Not too bad for a technical term!

It may seem odd that in a memoir about Ben Wright there are so many references to my publications and so few (only one!) to Ben's. In fact, this reflects the fact that most of what I learned from Ben was through personal interaction with him, and also that he has had such a strong influence on my academic career.

[3] I had to use the search terms "'Wright Map' measurement", as 'Wright Map' on its own resulted in lots of references to Frank Lloyd Wright.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Draney, K., & Wilson, M. (2009). Selecting cut scores with a composite of item types: The Construct Mapping procedure. In E. V. Smith & G. E. Stone (Eds.), *Criterion-referenced testing: Practice analysis to score reporting using Rasch measurement* (pp. 276–293). Maple Grove, MN: JAM Press.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessments*. Committee on the Foundations of Assessment. Pelligrino, J., Chudowsky, N., and Glaser, R. (Eds.). Board on Testing and Assessment, Center for Education. Division on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Wilson, M. (1984). *A psychometric model of hierarchical development*. Unpublished doctoral dissertation, University of Chicago.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*(2), 276–289.

Wilson, M. (2001, October). *On choosing a model for measuring*. Invited paper at the International Conference on Objective Measurement 3, Chicago, IL.

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research, 8*(3), 1–22. Download: http://www.dgps.de/fachgruppen/methoden/mpr-online/ (Reprinted in: Smith, E.V., and Smith, R. M. (Eds.) (2004). *Introduction to Rasch Measurement*. Maple Grove, MN: JAM Press.)

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice: Hypothesized links between dimensions of the outcome progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science*. Rotterdam, The Netherlands: Sense Publishers.

Wilson, M., & Draney, K. (2000, May). *Standard Mapping: A technique for setting standards and maintaining them over time*. Paper in an invited symposium entitled "Models and analyses for combining and calibrating items of different types over time" at the International Conference on Measurement and Multivariate Analysis, Banff, Canada.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12–14, 2000)* (pp. 325–332). Tokyo: Springer.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

Wright, B. D., Enos, M. M., Enos, M., & Linacre, J. M. (2001). *Adventures in questionnaire design: Poetics, posters, and provocations*. Chicago: MESA Press.

Chapter 9
Ben Wright's Kinesthetic Ventures

Ed Bouchard

Abstract In 1990, at the age of 64, Ben Wright came to me for a series of Alexander Technique (AT) lessons. This work eventually resulted in our collaboration on a book documenting how the internal subjective processes of physical experience coalesce in objectively reproducible ways across individuals (Bouchard & Wright, 1997). Ben initially sought out an AT practitioner at the suggestion of one of his friends from his childhood years at the Little Red Schoolhouse, Ann Sickles Mathews, an AT teacher in New York City. Then, in 2006, a member of Ben's family asked me to go over papers from Ben's office and to interview him on a near daily basis to begin to construct his biography. The purpose was to give Ben something he could work on, since his 2001 brain injury precluded future work in science. So, over the course of three years, I had the opportunity to begin sorting the through wealth of information in Ben's papers, and to record his answers to my questions about his life. What follows is a summary of some of that material.

9.1 Introduction: The V-12 Midshipman

One mystery that emerged early in conversations with Ben concerned his references to being in the Navy during WWII, and at other times to being at Cornell. How could he have been in both at once? In June of 1944, WWII raged on. The United States military had a shortage of scientists—and needed persons adept at math. For newly enlisted and drafted persons who were especially talented at math, the military assigned them to college classes at universities where the senior scientists were engaged in the war effort.

E. Bouchard (✉)
Independent scholar, Chicago Center for the Alexander Technique, Director, Chicago, IL, USA
e-mail: ed@edbouchard.com

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_9

Military recruiters came to The Hill School, the Pennsylvania boarding school Ben attended, and encouraged him to enlist in the Navy. So, at age 18, upon finishing high school, Ben signed up.

Before entering, he took the Army-Navy College Qualification Test (ANCQT). In light of Ben's later interactions with Educational Testing Service (ETS), it was ironic that this test was administered by none other than Henry Chauncey, a co-founder of ETS (Lemann, 2000, pp. 55 ff.). The test was an adaptation of a test first designed by Carl Brigham on the model of the Stanford-Binet IQ test. The ANCQT was further adapted to military needs by John Stalnaker and became the model for the ETS administration of the SAT.

Ben was assigned to the US Navy V-12 program for Midshipmen, which lasted from 1943 to 1946; it included 125,000 college-aged recruits, whom the Navy assigned to 131 colleges and universities throughout the United States. Ben's score on the ANCQT was such that the Navy assigned him to the Cornell physics department. Thus, he spent his tour of duty studying physics and electrical engineering at Cornell University at Ithaca, New York.

The policy was that V-12 Midshipmen were on active duty and they were expected to wear uniforms and subject to strict military discipline. However, at Cornell, the dress code was seldom enforced. In Ben's experience, military duty at Cornell felt like attending college at just about any time. Nevertheless, V-12 Midshipmen were required to carry a heavy 17 credits per college term; additionally, they were to do nine and one-half hours of physical training per week. For that Ben joined the Cornell swim team, in which he lettered.

Ben's Cornell transcript refers to 87 credit-hours in electrical engineering, in an era when computer science was a sub-domain of electrical engineering. A clue to his specific duties (and the content of his studies) can be found by examining the Cornell physics department roster of faculty, which included Hans Bethe, group leader of the Theoretical Physics Division at the Los Alamos Scientific Laboratory—and, Richard Feynman, group leader of the Los Alamos Calculations Division, who joined the Cornell faculty in the fall of 1945. With John von Neumann and colleagues, Feynman's Los Alamos special project was adapting an IBM business machine to solve the physicists' computationally intensive linear algebra equations. This by-product of the Los Alamos research was a step toward developing the modern computer (Rall, 2006).

Events outside of physics had a profound impact on Ben's science—and life. He enjoyed military discipline and might have stayed in the Navy. His uncle and a first cousin on his father's side were West Point graduates. Ben experienced health issues over the course of his life that may have been significantly mitigated by the habit of daily swims he initiated while a member of the Cornell swim team. It seems that Ben's favorite teacher at Cornell was neither Feynman nor Bethe but his swim coach, Scotty Little. With Little's instruction, Ben honed his backstroke into championship form. After an honorable discharge from the Navy on June 15, 1946, he continued for many years to swim a mile a day first thing in the morning.

9.2 From Bell Labs to the University of Chicago

In June of 1947, after graduating from Cornell in less than 3 years with distinction in physics and electrical engineering, Ben took a position at the Bell Telephone Laboratories under the direction of Charles H. Townes. During WWII, Townes designed radar systems for the Navy. After the war, Townes turned attention to the structure of molecules and characteristics of nuclei within atoms, investigations that eventually led Townes to develop the laser.

Ben's tenure with Townes was a short-lived summer between semesters, which suggests an unpaid internship rather than employment, lasting only from June to September, 1947. For the project, Ben used a 16-foot vacuum tube to test the microwave absorption spectra of the iodine monochloride molecule. He ran the experiment, collected the data, analyzed it, and wrote up a report. Ben's immediate supervisor, F. R. Merritt, offered it for publication in *Physical Review* (Townes, et al, 1948)—listing their young intern Ben Wright as third co-author. It was Ben's first scientific publication.

The purpose of the study was to map the separate contribution of the electrons associated with iodine from the electrons associated with chlorine, employing a weighted matrix factor analysis. While units such as angular momentum and Planck's constant are different, the structure of the results foreshadowed Ben's later work in social science.

In the fall of 1947, Ben accepted a Fellowship to continue investigations into molecular structure, but now under the direction of Robert S. Mulliken (1896–1986) at the University of Chicago Laboratory of Molecular Structure and Spectra. With Neils Bohr and colleagues, Mulliken changed physics. As he had done in his earlier research with Townes and Merrick, Ben's research goal with Mulliken and his colleagues was to designate the contribution of each of the individual elements. In Ben's iodine monochloride study with Townes, the purpose was to identify the contribution of two elements: electrons associated with iodine and electrons associated with chlorine. Research with Mulliken became vastly more complex; they might use an array of vacuum tubes, modeling the dizzying complexity of how each of the electrons related to all of the others as they mutually bonded or repelled.

For the molecular orbital spectroscopists, the goal was to construct a dynamic model of the specific contribution of various electrons inside the molecule interacting with the whole structure. It was daunting. As physicists, the first thing they did was to apply what they knew how to do, which was matrix linear algebra. Not being able to see inside a molecule, they made mathematical inferences from the data at hand, akin to modeling a transparent cubic structure in which the parts are constantly moving, and rotating the structure to get different views on how the parts interact. Their math envisioned the whole of the changing structure by, at the most basic level, constructing a map of the contributions of each of the various elements. This work foreshadowed Ben's later work in psychometrics even more than his previous studies with Townes.

Quantum mechanics is computationally intensive. Before computers, it was nearly impossible to do. Ben's mentors, Feynman, Townes, and Mulliken (and Mulliken's University of Chicago colleagues Clem Roothaan and John Platt), were pioneers in adapting computer language to perform their extensive calculations. Nevertheless, like a journey of a thousand miles, each step began with the simple arithmetical equivalent of putting one foot in front of another. Ben's training in electrical engineering with Feynman at Cornell served him well. In fact, to his surprise, now age 22, quantum electrodynamics was coming to seem too easy. He claimed to be finding "physics kind of boring," lamenting that measuring is "all [that] physicists do in their whole careers," telling himself he wanted "a livelier life" (Wright, 1988).

Ben had considerable respect for each of his colleagues. Mulliken, who was adept at making inferences directly from quantitative data, never prepared his lectures. For Ben, they were examples of thinking in action. His personal relationships with Mulliken and his University of Chicago colleagues were satisfying. Ben had become close friends with his lab partner Clem Roothaan, who joined on as a graduate student in the lab the same year Ben did. He liked his supervisor John Platt, too. His connections with Mulliken, Roothaan, and Platt were similar to the strong bond he had established with Townes. However, he missed Townes's profound envisioning of the unity between art, religion and science. Townes, who would eventually win both the Nobel Prize in Physics and the Templeton Prize in Religion, mused that

> …in our wider culture…we split apart the humanities and the sciences. Our cultural sensibilities separate…humanistic affections from science like we separate warm from cold, the poem from the microscope, the living from the dead. We sometimes even separate nature from science, assuming that nature is warm and that this warmth can be apprehended only by the poet. The scientist, in contrast, allegedly invokes cold-hearted methods that are as deadening to nature as they are to the human spirit….[We] assume [there] is a split between the aesthetic and the empirical, a rift between human affections and disciplined research. But a closer look will show that *aesthetic affections are alive and well in the scientist*. The vastness of the apparently limitless reaches of outer space combined with the intriguing complexities of the smallest microscopic things elicit an aesthetic response in the human soul. Nature communicates meaning; and science can actually facilitate this communication. (Townes, 2001, pp. 297–8; emphasis added)

From investigations into molecular structure with Townes and Mulliken, Ben gained exciting firsthand experience in science. From Townes he had an inspirational vision of uniting science, the arts, and spirituality. From Mulliken he had powerful examples of on-the-spot quantitative modeling. Now, at the University of Chicago, free of Navy requirements, he found aesthetic affections were indeed alive and well within him. Ben began taking English and social science classes; he studied psychology with Carl Rogers, and sociology with Lloyd Warner (Raines, 2002, p. 212). If "nature communicates meaning" and "science can actually facilitate this communication" as Townes taught, Ben wanted to bring out this unity and tell others about it too. Ben was discovering that poetry and physics come from the same place inside and that a successful theory of science will have the same properties as a good poem, and vice versa.

Ben demonstrated the mathematical facility needed for navigating quantum electronics alongside the world's top physicists, including future Nobel Laureates Bethe, Feynman, Townes, and Mulliken—and was fully committed to science. Yet, he yearned for something else. More and more often he found himself feeling trapped. There were almost no social interactions in a physics laboratory.

In his first semester, he found some respite directing a group theatre for young adults at the Gads Hill Center, a settlement house in Chicago. The contrast between the theatre world and physics impressed him. When he observed his physicist colleagues, it seemed to him that they suffered from what Townes had described as "cold-hearted methods…as deadening to nature as they are to the human spirit."

Casting about for an alternative path, Ben attended several lectures of the psychometric pioneer, Louis Thurstone, a co-founder of the Psychometric Society and the journal, *Psychometrika*. He quickly adjudged that while Thurstone asked the right questions to construct social science measures, factor analysis was an incomplete way of answering them. Later, Ben would see that Thurstone had made a mistake in abandoning his rigorous mathematical scaling principles in favor of factor analysis (for Thurstone's account, see Thurstone, 1952, pp. 311–312; Thurstone, 1959, pp. 214, 321; also see Lumsden, 1980, p. 7). Ben wondered whether he could make a contribution to social science. He later recalled,

> It was a brilliant Spring morning. The birds were chirping. The girls and boys were flirting and I was copying giant quantum mechanics equations from Willi Zachariasen's blackboard. So I put down my pencil, left the class and left Physics. I worked as a laboratory physicist for a few more years, because I needed the money and had the skills, but I went in search of life. (Wright, 1988, p. 25)

Ben's phrase "went in search of life" suggests another influence driving his career choice. In Second Grade and again from Fourth to Seventh Grade, Ben attended the Little Red School House in Greenwich Village, a private elementary school in Manhattan, founded by education reformer Elisabeth Irwin (1880–1942). Irwin was an articulate advocate for teaching "the whole child." With respect to Ben's remark that he was going "in search of life," consider what Irwin wrote in 1924:

> [W]hat is education all about? It is not primarily a process of imparting information; it is not first of all a method of teaching reading and writing and thereby ridding the country of illiteracy. It is to provide situations in which a child can *experiment with life*, can express himself creatively, can orient himself in his own world. In fact the school is or should be a place where a child's physical and emotional energy may be released for his own purposes, where he can learn to act on his own initiative and take the consequences of his own acts. If a school can provide these conditions, the young individual will then gain a sense of power over his environment. (Irwin, 1924, p. 8; emphasis added)

And in 1928 Irwin wrote:

> Modern psychologists and mental hygienists tell us that those people are happiest and healthiest who can best adjust to reality, can meet life face to face. The school then, if it is to help individuals to be efficient and active members of society, *must introduce children into life* rather than shelter them from it. It must be a laboratory rather than a monastery.

> The task of education today—to change our schools from monasteries into laboratories, laboratories not where educators experiment with children but where *children experiment with life*. (Irwin, 1928, p. 273; emphasis added)

In the 1990s, Ben referred to his years at Little Red as his best and most influential formal educational experiences. One teacher, Florence Beaman, was especially important to Ben. To enhance his interactions with his age peers, Miss Beaman taught Ben to look for general patterns in how other people fit into the social fabric of his or her life. Miss Beaman understood that seeing patterns is what Ben did naturally. When Ben later understood that seeing patterns was algebra, he realized an important aspect of substantive mathematical thinking as it applied to experimenting with life. Losing his connection with that early lesson in graduate school pushed Ben to look in new directions.

9.3 An Apparent Decision for Psychology over Physics

With a career track in physics behind him, Ben chose a bumpy road. In 1950, he signed on as a Counselor at the University of Chicago Orthogenic School. Ben formed a close friendship there with Carl Rogers, himself a maverick. Roger's soft-spoken manner seems more personally appealing than that of the autocratic Bettelheim. Rogers, however, did not have two things Bettelheim did: a personal example of survival against all odds and a school that was based on the inspiration that drove Little Red. Milieu therapy at Orthogenic School and experiential learning at Little Red stemmed from like experiential learning theory principles.

Under the direction of Bruno Bettelheim, the Orthogenic School was a residential school for school aged children diagnosed with mental illness. Bettelheim often went out of his way to accept children that other institutions had rejected as too difficult for a therapeutic setting (Raines, 2002, p. 231). Although Bettelheim was known for his stern demeanor, Ben found him warm and good-hearted. He admired the way he ran the school. Bettelheim reminded Ben of both Miss Beaman and his beloved grandfather, though they were each personally quite different from one another.

While also completing his course work and continuing to work in the physics lab, Ben earned a Certificate in Psychoanalytic Childcare from the Chicago Institute for Psychoanalysis (1954) en route to a Doctor in Philosophy of Human Development from the University of Chicago (1957), and an Illinois State license to practice clinical psychology (1959, 1964).

For most, psychoanalytic psychology seems a far cry from quantum physics. For Ben, a psychoanalytic interaction begins with collecting data. Observations and stories from a patient's life accumulate into a tangle of raw data that are not so different from data generated by the spectra of a physics lab. From this tangle, a theory emerges that can model forces invisible to the naked eye, forces driving the life of the person and the situations they are in. Finally, an hypothesis as to an effective

intervention can be formulated, and further observations can be made to test the hypothesis, modify the theory, revise the intervention, etc.

Psychologists employ various social science research designs, including interviewing, taking life histories, writing reports, constructing tests to bring out "unconscious" material. At the Orthogenic School, Ben did all of that as well as charting the treatments given, the counselors' observations and their assessment of the progress of the children. In an interview with Ben about his experience at the School, Bettelheim biographer Theron Raines (2002) found that the methodology that most resonated with Ben was the simple taking of extensive notes of one's observations.

> [Ben] made the point that the counselors' reports were a brilliant teaching device. Bettelheim read them all, thousands of pages a year, and where he saw something crucial happening, he marked the margin to guide or jolt the staffer.
>
> "The counselors had to think about what they did," Ben said, "and then Bruno thought about what they *thought* they did, and then they thought about what *he* thought."
>
> [I]nducing reflection in this way [Ben observed] led counselors to dig for deeper meanings and to remind themselves of their goals as their understanding grew…
>
> "If you just go on," [Ben] said, "and don't think about what you're doing, you bounce from one impulse to the next, you keep paddling so you don't sink, but you don't know where you're going. [Writing the reports] forced them to think, which gave them a second voice. That's where culture comes from. That's what intellect is about." (pp. 232–233)

In his subsequent career as a teacher, Ben used such report writing—and his own responding to them—as one of the cores methodologies for teaching and for generating new ideas in science. Bettelheim, however, resisted conducting research into the effectiveness of the treatment of the Orthogenic School. Ben took an opposite view, and felt that if he continued on the clinical psychology career path, he would be as unhappy as he had been in physics. Bettelheim twice offered Ben the directorship of the Orthogenic School. Ben turned him down both times, and Bettelheim stayed out of Ben's way, allowing him to pursue his own course. As Townes had given Ben a vision for a unity of art, science, and religious inspiration, and as Mulliken taught him advanced techniques for inferring meaning from patterns in numbers (Mulliken, 1972), Bettelheim refined Ben's capacity for seeing how teachers and their teaching are shaped by the experience of interacting with students. Two publications (Bettelheim & Wright, 1955; Wright & Bettelheim, 1957) document this lesson and provide yet another foreshadowing of Ben's later contributions in psychometrics. The path in that direction opened up when the University of Chicago Education Department encountered a sudden need for an instructor in introductory statistics. Ben was offered the position, purportedly because of his ease around numbers (Linacre, 1998).

Ben started teaching statistics in 1956, but soon ran into trouble. He noticed that the statistical textbook incorporated multiple errors in its recommendations and processes. Ben followed his training in physics, and started teaching statistics as he understood it rather than what was in the textbook. This soon drew the ire of the Education Department faculty as they encountered students unfamiliar with the expected statistical methods. The Chair of the Department, Frank Chase, supported Ben, but the controversy grew to the point that the University's foremost statistician,

L. J. Savage (1917–1971) was consulted. Savage also supported Ben, consolidating his reputation as a knowledgeable and reliability authority in quantitative methods (Linacre, 1998, pp. 23–24).

For Ben, teaching statistics and research design was a major step in merging the split between his interest in people and his passion for research. He relished challenging fundamental beliefs about statistics as much as he enjoyed the mathematics. Thurstone had sought to avoid controversy, abandoning his truly significant advances in measurement theory and practice in favor of factor analysis. Wright, in contrast, not only positioned himself to become a staunch advocate of such advances, but deftly used disagreements over principles in a way that enhanced the professional identity of his students (see the chapters by Linacre and Fisher, this volume).

9.4 Factoring in the Univac I

Another pair of milestones in Ben's journey occurred in 1959 when the University of Chicago received a gift of a Univac I (1 kB) vacuum tube computer, and then again in 1962, when the university received a $2.5 million IBM 7090 mainframe computer. The latter took up the entire basement of the Institute for Computer Research at 5640S. Ellis. A computer was a tool unfamiliar to social scientists. Ben's experience translating quantum electrodynamics into linear programs prepared him for the opportunity to put the Univac I to use. He promptly wrote a program to perform factor analysis.

Introduced by British statistician Charles Spearman in 1904, factor analysis had been further developed by University of Chicago psychometrician Louis Thurstone in the 1920s. It requires performing multiple correlational computations on a (usually large) data set to reveal underlying patterns in the data. Factor analysis became a core methodology in various sciences, growing far beyond its specific application in intelligence testing.

Access to computers gave Ben a remarkable advantage. Previously, a single factor analysis could take months to complete. Ben (1988) noted that even founders of factor analysis like Spearman and Thurstone had performed only maybe 20 or 30 in their lifetimes. Now, even with the early computers of the late 1950s, Ben was able to churn out that many per week, hundreds over just a few years.

Applied research helped Ben support his growing family. Ben worked as a consultant to Social Research, Inc. (SRI), led by Burleigh Gardner. His supervisor was Lloyd Warner (1898–1970), an anthropologist and sociologist noted for applying quantitative methods to contemporary issues. In the 1970s, Mark Stone, with whom Ben would write the first introductory and still widely read text on Rasch measurement (Wright & Stone, 1979), became a full partner in synthesizing an integrated perspective on models, estimation, and construct validation. In the early days of this work, Ben was distressed that the results of each factor analysis were sample- and analyst-dependent. When each new sample of the 'same' data yielded a different

factor structure, core repeatability was undermined. Finally, reluctantly, Ben concluded that something in factor analysis itself produced unstable results (Wright, 1988, 1996a). Ben suspected a problem with the mathematical underpinnings themselves. He was earning a living doing it, but increasingly felt factor analysis was not a viable path to scientific progress.

Given his track record, Ben's next career move should have been no surprise. He had left future Nobel Laureate Charles Townes at Bell Labs for the University of Chicago. He had abandoned a position as a research physicist with future Nobel Laureate Robert Mulliken to accept a tough job as a counselor in a residential center of children diagnosed as autistic and schizophrenic. Now, he left behind an opportunity to be one of the earliest to adapt computers to statistical analyses at the core of social science research because he was dissatisfied with the scientific quality of the results.

9.5 The Step

Even before the departmental hubbub over his critique of errors in statistical textbooks, Jimmie Savage and Ben had become friends. Savage had been a student of John von Neumann's at Princeton in 1943–1944 (Wallis, 1981, p. 15). He published a series of papers with the economist Milton Friedman critical of Keynesian economics; their names remain linked in ongoing references to the "Friedman-Savage utility function." Like Ben, Savage was a critic of classical statistics. In 1960, Savage invited a colleague he had met several years before when they both worked with the Cowles Commission on Economic Research, the Danish mathematician Georg Rasch, to lecture at the University of Chicago.

We will not go over that history here. Ben and others have told that story before (Fisher, 2008; Wright, 1988, 1992, 1996b, 2005). It may be, though, that, structurally, Ben's approach to analysis in social science had not changed much from his research as a molecular orbital spectroscopist. We can put Ben's career in perspective by considering a quotation from Charles Sanders Peirce:

> The scientific specialists—pendulum swingers and the like—are doing great and useful work, each one very little, but altogether something vast. But the higher places in science in the coming years are for those who succeed in adapting the methods of one science to the investigations of another. (Peirce, 1989, p. 380)

That is an apt summary of what Ben accomplished over the course of his career, repeatedly, applying natural science methods in social science, in applying psychology in his professional relationships, in seeing interactions with students as key to teachers' senses of themselves as teachers, and in seeing the construct clarifications of instrument calibration as an essential aspect of educational measurement.

In his studies with me learning the Alexander Technique, Ben saw that, though the act of balancing seems entirely subjective, we are usually simply unaware of the objective processes through which we constantly maintain physical positions

and movements. He understood that we constantly make intuitive ordinal judgments about how much force—physical, social, illocutionary, or otherwise—to apply in various situations as we seek to balance ourselves relative to others and things in the world. What Ben has done through his psychometric research is to show that by framing and posing the right questions about these, and all, ordinal judgments, we can obtain meaningful quantitative measures that enable us to be more fully and truly what we are. How we will put this discovery to use is now in our hands. For this, psychology and the social sciences are indebted to Ben Wright.

References

Bettelheim, B., & Wright, B. D. (1955). Staff Development in a Treatment Institution. *The American Journal of Orthopsychiatry, 25*(4), 705–719.

Bouchard, E., & Wright, B. D. (1997). In M. Protzel (Ed.), *Kinesthetic Ventures: Informed by the work of F.M. Alexander. Stanislavski, Peirce & Freud*. Chicago: MESA Press.

Fisher, W. P., Jr. (2008, March 28). *Rasch, Frisch, two Fishers and the prehistory of the Separability Theorem. In Session 67.056. Reading Rasch Closely: The History and Future of Measurement*. American Educational Research Association, Rasch Measurement SIG, New York University, New York City.

Irwin, E. (1924). *Personal education*. The New Republic, Educational Section, 40(519:II), 7–9.

Irwin, E. (1928). We watch them grow. In *Survey Associates, Charity Organization Society of the City of New York* (Vol. (60)). New York: Survey Associates.

Lemann, N. (2000). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus and Giroux.

Linacre, J. M. (1998). Ben Wright: The measure of the man. *Popular Measurement, 1*, 23–25.

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement, 4*(1), 1–7.

Mulliken, R. S. (1972). Spectroscopy, molecular orbitals, and chemical bonding (Nobel Lecture, December 12, 1966). In *Nobel lectures, Chemistry 1963-1970* (pp. 131–160). Amsterdam: Elsevier.

Peirce, C. S. (1989). Lecture on logic, 1882. In C. J. W. Kloesel, M. H. Fisch, N. Houser, U. Niklas, M. Simon, D. D. Roberts, & A. Houser (Eds.), *Writings of Charles S. Peirce: A Cronological Edition: Volume 4 1879-1884*. Bloomington/Indianapolis: Indiana University Press.

Raines, T. (2002). *Rising to the light: A portrait of Bruno Bettelheim*. New York: Alfred A. Knopf.

Rall, D. N. (2006). The 'house that Dick built': Constructing the team that built the bomb. *Social Studies of Science*, *36*(6), 943–957.

Thurstone, L. L. (1952). L. L. Thurstone. In G. Lindzey (Ed.), *A history of psychology in autobiography* (Vol. VI, pp. 294–321). Englewood Cliffs: Prentice Hall.

Thurstone, L. L. (1959). *The measurement of values*. Chicago: University of Chicago Press, Midway Reprint Series.

Townes, C. (2002). *How the laser happened: Adventures of a scientist*. St. Louis: San Val.

Townes, C. H. (2001). Logic and uncertainties in science and religion. In *Science and the future of mankind: Science for man and man for science* (pp. 296–309). Vatican City: Pontificia Academia Scientiarum.

Townes, C. H., Merritt, F. R., & Wright, B. D. (1948). The pure rotational spectrum of ICL. *Physical Review, 73*, 1334–1337.

Wallis, W. A. (1981). Tribute. In *The writings of Leonard Jimmie Savage: Memorial service tributes* (pp. 11–24). Washington, DC: The American Statistical Association and the Institute of Mathematical Statistics.

Wright, B. D. (1988). Georg Rasch and measurement. *Rasch Measurement Transactions, 2*(3), 25–32.
Wright, B. D. (1992). The International Objective Measurement Workshops: Past and future. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1, pp. 9–28). Norwood: Ablex Publishing.
Wright, B. D. (1996a). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*(1), 3–24.
Wright, B. D. (1996b). Key events in Rasch measurement history in America, Britain and Australia (1960-1980). Rasch Measurement. *Transactions, 10*(2), 494–496.
Wright, B. D. (2005). Dedication: Memories from my life. In N. Bezruczko (Ed.), *Rasch measurement in health sciences* (pp. vi–xvii). Maple Grove: JAM Press.
Wright, B. D., & Bettelheim, B. (1957). Educational news and editorial comment: Professional Identity and personal rewards in teaching *The Elementary School Journal, 57*(March), 297–307.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

Chapter 10
Statistical Models, Scientific Method and Psychosocial Research

Raymond J. Adams

Abstract This piece is a compilation of a number of short class-notes I wrote in 1987 and 1988 as a result of discussions with Ben and fellow students whilst I was a student at the University of Chicago. At that time Ben was pushing us to consider why progress in the psychosocial sciences seemed to be so frustratingly meagre when compared to progress in the 'hard' sciences. In discussions with Ben it seemed, to me at least, that central to his argument was a view that much of 'so-called' statistical modelling was unscientific—that it focussed on the description of ad-hoc collections of existing data, rather than proposing and rigorously testing of models and theories through the analysis of measures with well understood properties. Ben was very critical of exploratory statistical analysis, made a clear distinction between measurement models and analytic models and was always reluctant to fit statistical models to data—he wanted to use statistics as a tool to test whether data were consistent with theoretically posited models, he wanted to fit data to models. One wonders how he would have felt about the current big data and data mining movements.

10.1 Note

The material below is not meant to be profound, nor should it be read as presenting a well-developed view on statistical modelling. What I hope it does is give an insight into the nature of the discussions Ben held with his students and in class during those days.

10.2 Psychosocial Research Methods

Psychosocial research involves the study of human behaviour and social phenomena. Depending upon the context it has been referred to as 'social science,' 'human science,' 'social inquiry' or 'behavioural science.' These labels are

R.J. Adams (✉)
ACER, Melbourne, VIC, Australia
e-mail: ray.adams@acer.edu.au

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_10

generally meant to encompass a core of disciplines such as Sociology, Psychology and Economics, and fields of application such as Education and Social Welfare.

The role of scientific method, as an approach to understanding the human world, has become an ongoing issue of debate in psychosocial research. Indeed, the philosophy of science is filled with debate about the validity of scientific method (in any context) (see for example Chalmers, 1976; Polkinghorne, 1983; Phillips, 1987). Despite the debate, it is recognised that scientific method has been successful in its application to the study of the natural world and, recognizing this success, workers in psychosocial research have attempted to apply scientific method to the human world.

In this paper, I do not intend to produce long and involved arguments regarding the definition of scientific method, as it is applied to psychosocial research, it will be useful for later discussion to present a simple framework for describing science. The framework that is presented is a personal one but draws upon Fowler (1962), Chalmers (1976), Polkinghorne (1983) and Phillips (1987).

I will then mention some of the problems that have been encountered in the application of scientific methods to the human world and in the final section I will discuss the application of statistical models to psychosocial research. The application of these methods may play a role in the disappointing outcomes of scientific enquiry in psychosocial research.

10.3 What Is Science?

Science is the dominant process employed by man to achieve a knowledge of the world. Science is driven by a view that knowledge can be expressed as a set of publicly accepted and infallible truths. Science never claims to have actually found these truths but it does claim to be getting closer to them. This identifies an important fallacy regarding science that should be dispelled. “Science” does not mean “truth.” Hence, the term “scientifically proven” as is claimed by so many television commercials, is a logically inconsistent statement.

Science, like all desires to understand the world, is motivated by a need to survive. All areas of man’s inquiry develop from investigations originated to increase chances of survival. In its motivation, science is no different than any other method proposed as a tool to help us understand our world.

There are no definitive characteristics of science but it can be distinguished from other methods of inquiry by the principles of knowledge on which it is built. The first is the principle of objectivity, public truth and verifiability. For knowledge to be scientific it must be possible for that knowledge to be independently verified. At this point we are walking a fine line, since “objectivity, public truth and verifiability” could easily be construed as requiring absolute truth.

This is not however the case–science does not require knowledge to be absolute; what it requires is the possibility of verification and publication, within the constraints of a given, shared set of values and a common theoretical stance. A second principle is that the outcomes of science must be useful. This does not

mean that science must always fulfil a specified need, often the practical uses of scientific discoveries are not identified until well after a discovery has been made. In fact, there is a strong argument to be made for science as a creator rather than satisfier of needs. Science is useful when it leads to a simpler yet more comprehensive understanding of the world. Eisner (1979) combines the principles of objectivity as follows.

> … objectivity is a function of intersubjective agreement among a community of believers. What we can productively ask of a set of ideas is not whether it is really true but whether it is useful, whether it allows one to do one's work more effectively. (Eisner, 1979, p. 214)

The third principle for science is that our current knowledge is incomplete and represents only a construction that attempts to explain reality and that construction must always be open to development, modification or rejection.

The process of science is one of making observations and making inductions from those observations to develop a "theory." The method of observation must always be consistent with the principle of objectivity and the process must be ongoing in recognition of the principle of incomplete knowledge. Kinston (1985) spells out five stages that are involved in scientific work. I will take some time to describe each of Kinston's stages since they are important in the following discussion.

Kinston's first stage, level I, is "entity." All knowledge begins with the formation of ideas. In examining reality, the scientist begins by creating an idea or concept. These concepts can be very general. For example, they could be anything from "ability" to "heat" or even quantity. The entity is subjectively defined.

Level II, "observable," involves two ideas. We must take our original idea and add "thingness" so that our original idea or concept can become public. In practice, Level II involves the definitions of conditions and criteria that enable the original concept to be operationalized.

At Level III, "comparable," the concept of quantity is added to "thingness" and our original idea. According to Kinston "a comparable is formed by ordering and ranking observables and answers the question: 'which is more (less)' or 'which is better (worse)?'" (p. 98). According to Kinston, Level III requires the subjective use of the concept of quality.

For Level IV, "measurable," we add the idea of "generally applicable unit." Quantity is taken for granted and we establish a measurable that enables us to describe "how much" in an absolute sense. Clearly, that absolute must be defined with respect to some standard. Kinston sees the move from levels I to II and from III to IV as a process of objectifying the idea. This use of objectivity corresponds to the one used as a principle of science. That is the "idea" is becoming more public, within a specified set of constraints and conditions. Note that subjectivity also plays a role in science. This subjectivity plays a role at level I, where the original idea is private, and at level III where the subjective sense of quantity is introduced. We attempt to impose objectivity by moving from levels I to II, and from levels III to IV.

The last of Kinston's levels is Level V, "relatable." The idea of "relation" is added to our previous four and we begin to describe the world in terms of relationships between the subjective ideas we began with.

According to Kinston then, the result of scientific inquiry is the development of "theories" that specify relationships between measurables in "a deliberate attempt to model or represent significant aspects of reality" (p. 95). We could add that these theories will support predictions and explanations of events within a defined class. The boundaries of that class are specified by the contents and specifications used in the development of the entities, observables, comparables, measurables and relatables that make up the theory. The principle of incomplete knowledge warns that scientific theories are not infallible and they must always be open to modification. The scientist must always be prepared to go back and modify the construction at any level to improve the utility of the theory.

While there is no way for scientific theories to be proven correct they must always be open to refutation. In fact, Popper sees testability and openness to refutation as the essence of scientific inquiry (Phillips, 1987).

10.4 Science in Psychosocial Research

The effectiveness of scientific methods of inquiry in the natural sciences has led to its adoption as a paradigm for psychosocial research. But even the most ardent proponents of scientific methods in the human sciences have recognized that the achievements thus far have been a little disappointing. For example, Hedges (1987) comments that:

> Psychologists and other social scientists have often compared their fields to the natural (the "hard") sciences with a tinge of dismay. Those of us in the social and behavioral sciences know intuitively that there is something "softer" and less cumulative about our research results that about those in the physical sciences' (Hedges, 1987, p. 443)

Many factors have been identified as possible causes for the apparent failure of scientific methodology in the psychosocial sciences—Hedges (1987) lists a number of references that discuss possible explanations for the perceived failure and limitations of scientific methodology in psychosocial research. Valentine (1982) emphasizes two possibilities. The first is a lack of systematicity. She believes that science relies on systematicity in the subject matter so that a coherent body of knowledge can be developed and that a lack of systematicity in the human world causes problems in the definitions of variables that are suitable for the expression in a coherent body of knowledge. The second is generality. She claims that scientific theories are unrestricted by space and time, an ideal that cannot be met in research on the human world. The arguments against the suitability of scientific method in psychosocial research can be persuasive and many are not without merit.

The application of scientific method to the human world may well be more difficult than the application of scientific method to the natural world. But when identifying sources of failure for a particular research paradigm we should not only examine the subject matter and its suitability for us with the paradigm, but we should also examine the fidelity with which the paradigm was employed. Before we begin to criticize the appropriateness of the scientific method in psychosocial

(because of its apparent failure) perhaps we should examine the fidelity with which scientific method has been employed.

Some of the most common misconceptions in psychosocial research have centered around the use of quantitative data, experiments and sophisticated statistical models. It would not be unfair to argue that most researchers believe that the more of these three factors you have, the more scientific your study is. Perhaps this is part of the problem of scientific method in psychosocial research. Quantitative data, experiments and statistical models do not make science.

In the remainder of this paper, I will examine the case of statistical models and comment on when their use may be scientific and when it may not.

10.5 Models

The term *model* has widespread use throughout all research. We have for example: The general linear model, models for pattern recognition, internal models of attachment figures, computer models of learning processes, stochastic models for learning and Rasch models, to name just a few. In each case a model acts as a representation of reality. Models have proven fundamental in all forms of inquiry and they have a central role in scientific method.

In most cases models are expressions of theories but the form of the expression will depend on the purpose of the models. First, a model can be used to assist in the *explanation of theory*. This is usually done by constructing the model with terms, concepts and images that are more readily understandable than the theory itself. An important purpose of models lies in the *testing of theory*. When formulated as a model, logical inconsistencies in the theory may be identified. In some form, models can be tested through simulations of reality and, when expressed in particular mathematical forms, a range of statistical methods are available to "test" the theory. Through improved explanation of theory and testing of theory the models can them lead for further development and enhancement of theory.

One of the largest classes of models used in psychosocial research are the statistical models—or the "off-the-shelf" variety of statistical methods. Note that those mathematical and statistical models that were developed for a specific purpose or research situation and that do not enter common usage are not meant to be covered by this discussion.

It is not an uncommon view amongst social scientists that the use of these models leads to a scientific research study. But to what extent do these methods act adequately as models? What role do these methods play in the scientific process described above? Based on these considerations, when are these models applied scientifically?

Table 10.1 A classification of statistical methods

Name	Example methods
Descriptive methods	Exploratory factor analysis
	Descriptive statistics
Explanatory methods	Log-linear modelling
	Stepwise regression
Confirmatory methods	Linear structural relations
	ANOVA
	Confirmatory factor analysis
Axiomatic methods	Rasch measurement

10.6 Categories of Statistical Models

To discuss their application to psychosocial research it is useful to construct a fourfold classification of statistical models. One possible classification scheme is presented in Table 10.1. The allocation of an approach to a category may depend on the mathematical form of the model, but it is more likely to depend on the methodological reasoning underlying the use of the method. For example, factor analysis, depending upon the details of its application may be classified as an exploratory or confirmatory approach. Each of the methods can be discussed in terms of the relative role of: the data, substantive theory and the constraints imposed by the mathematical form of the model.

10.6.1 Descriptive Methods

These are used to "fish around" in data. When the researcher has a body of data that has been observed and has no theory, it is possible to use mathematical methods to manipulate the data in a search for relationships that may be useful in a development of the theory. Perhaps the two most common exploratory methods, beyond simpler descriptive statistics are correlation and factor analysis. In many instances when a research (data analyst) is faced with a body of numerical data for which he/she has no theory, a set of correlations will be calculated to identify any covariation between variables. Substantive theory is then built to explain the observed covariation. Exploratory factor analysis is a more systematic approach to the examination of correlations. The aim of factor analysis is "the resolution of a set of variables linearly in terms of a small number of 'factors'" (Harman, 1976, p. 4).

In these techniques, the appropriateness of the model for the data is rarely considered. Supposedly, the statistical techniques employed allow the patterns and relationships in the data to be exposed, while making only very weak constraints in that exploration. In descriptive methods, it is hoped that the analysis is driven by the data, with theory playing only a limited role through the mathematical specifications of the model. In some of the simpler descriptive techniques such as scatter

plots, histograms, box and whisker and the like, this assumption may be almost fulfilled. Beyond that things become less clear. Exploratory factor analysis and cluster analysis are obviously method bound and even the selection of a measure of central tendency (mean, median, mode) can have an impact on data interpretation.

10.6.2 Explanatory methods

These methods are used when developing models to describe a set of data. Their aim is to develop a mathematical model that accurately reproduces the observed data. Rather than being driven by theory these methods are driven by a combination of the data and the mathematical technology being employed. In general, the measure of success in applying these models is the degree to which the developed model confirms to the observed data (fit). In the development of these models there is always some tradeoff between model simplicity and the accuracy of the model in reproducing the data. The researcher must be careful to ensure that the plausibility, utility, elegance and simplicity of the associated theory does not get lost in the search for model to data fit.

In these explanatory methods, the form of the model places strong constraints on the development of substantive explanation. This is argued as valid on the basis that some models can generally be constructed to fit any given set of data. Unlike descriptive methods however, it is recognized that any structure identified, or developed from the data, is strongly bound by the researcher's approach to the analysis.

10.6.3 Confirmatory methods

Methods of this kind are used to test the plausibility of a theory when it is stated in a particular form. This category includes traditional approaches to experimental data analysis and the more recently developed confirmatory data analysis procedures. In both cases a mathematical model is constructed that is argued to be commensurate with the substantive theory to be tested. Mathematical and statistical techniques are then used to test the compatibility of the theory, as expressed by the model, with observed data. The most common approach in psych-social research is to take an 'off-the -shelf' statistical method and assume that it can be used to represent the substantive theory, then apply standard testing procedures designed for that method. The aim is to fit the model to the data. If the data does not fit the model, then the model is rejected and the theory (or the data collection method) is modified.

In the experimental case, the purpose is to test the plausibility of a specific hypothesis that has been proposed by the researcher. A mathematical model that is claimed to be commensurate with substantive theory is selected, and the model is then fitted to the data, and the acceptability of that fit is examined. In experimental

designs, the model is formulated so that rejection of the ("null hypothesis") model adds support to the researcher's theory.

These approaches are driven more strongly by theory than descriptive or explanatory methods. Theory is used to construct the form of the models and the theory, as represented by the model, is tested through fit to observed data. While the data is not allowed to "speak for itself" in the sense of descriptive and explanatory models, it is being used to test the possibility of a particular theory being true.

10.6.4 Axiomatic Methods

The mathematical models used with these methods are derived from a set of axioms required by the researcher. It has been argued that in some instances these mathematical models are deduced from the axioms. By definition, these axioms cannot be proven or disproven.

Examples of axiomatic methods are the application of Rasch models. Rasch models are developed from fundamental axioms regarding the desired or necessary nature of measurement. Given these axioms it is argued that if a measuring instrument is to be valid in the sense of having specific objectivity, then it must conform to an appropriate Rasch model. Specific objectivity means that, once calibrated, the data from any subset of fitting items may be used to measure a person, and vice-versa, that the data from any subset of fitting persons may be used to calibrate the items. When developing the measuring instrument, observations are made and an attempt is made to fit the observations to the model. If the data do not fit the model, then it is argued that the instrument does not provide a valid ("specifically-objective") measure. The researcher should then examine why his/her measurement intentions have not been met by the instrument that was constructed.

In this case the mathematical form of the model, which has been built upon a set of specifications (axioms), takes a dominant role. Theory and model are far more intimately related than in any of the other approaches. The theory and model may in fact be the same thing only expressed in different forms.

10.7 When Are Statistical Methods, Models?

Harré (1976) considers two types of models: *sentential* models and *iconic* models. A sentential model is a set of sentences in some kind of correspondence with another set of sentences. An iconic model is a thing, structure or process in some kind of correspondence with another thing, structure or process. Harre further adds that models whose subject and source differ have come to be called *paramorphs* and those whose subject and source are the same are called *homeomorphs*.

Just as paramorphs may be the subject matter of sentential models, so too may homeomorphs. The description of homeomorph may be treated as a sentential

model of the description of its source subject. I am inclined to think that this is the kind of modelling that we hope to do when applying statistical methods in psychosocial research. The sentences in the statistical method can be treated as a description of a homeomorph of the real psychosocial world.

The requirements of this kind of modelling include a correspondence between the sentences that make up the statistical model and the homeomorph that the researcher has constructed as a theory.

If we consider descriptive methods by these criteria we can see that they are not models at all. They are never intended to have any correspondence with a particular model of the psychosocial world.

Explanatory and confirmatory methods are both attempts at sentential descriptions of homeomorphic models and therefore do aspire to be legitimate models. Explanatory methods are attempts to build sentential models in the form of formulations of the relationships in observed data and confirmatory methods are both that and also attempts to test specific models against observed data. The validity of these models, for this purpose, depends upon their ability to reflect the researcher's homeomorph. Unfortunately, beyond the selection of variables for inclusion into the model this is rarely a major consideration of the practical researcher. The use of a method is often determined as much by its availability as its suitability for the problem at hand.

The axiomatic models belong more clearly to the class of sentential models. If we take the Rasch model as a particular example, then we can see that it forms a sentential model of the process of measurement.

In summary, we can see that descriptive methods are not models (and probably have no aspiration to be models). Explanatory and confirmatory methods need to be models if their application is to be valid; but they too often fall short in their correspondence with the researcher's other expressions of the same model. Finally, axiomatic methods are always models because they are built to be representations of specific theories.

10.8 What Role Do These Methods Play in Science?

The role of these methods can be further examined by looking at their place in the process of science as outlined by Kinston and discussed above.

All of the methods assume at least the first two of Kinston's levels—'entity' and 'observable.' Since all of the methods require the use of observations, these two levels must be first developed by the researcher. Although entity and observable must be defined before the application of statistical methods, one of their important uses is the provision of information for the modification and redevelopment of entities and observables.

In considering our four statistical methods it is our example of an axiomatic method, the Rasch model, that plays a unique role. The Rasch model is concerned with taking developments from the first three levels and constructing 'measurables' whereas descriptive, confirmatory and explanatory methods take observables and 'measurables' to produce 'relatables'.

Some explanatory, data fitting, methods are used in an attempt to construct measurables from comparables. The application of these methods to constructing measurables is however invalid. Since measurables can only be constructed by the addition of the entity 'generally applicable unit' to the comparable, the only valid statistical method is one that can govern the construction of that unit. An axiomatic model that is built specifically to take comparables and add generally applicable unit to provide us with measurables is necessary. The explanatory methods used for this purpose are not commensurate with the entity 'generally applicable unit' so they cannot be used to construct a measurable or test it.

While the range of descriptive, confirmatory and explanatory can be applied with data from the 'observable' or measurable levels, they are at their most powerful when they take measurables and examine the relationships between them to provide relatables. In some cases however, the nature of what we are studying may force us to search for relatables with observables or comparables.

10.9 When Are These Methods Applied Scientifically?

Each of these methods, if used wisely, has something to offer science although some are more likely to be of use than others. The Rasch model is a fundamental tool in making the construction of measurables possible and measurables are the most powerful variable that we can use in producing relatables. The confirmatory methods, if used with variables from the highest level feasible, and, when designed to be commensurate with substantive theory, are a powerful means of testing theory. But, if we do not ensure that the model matches the theory and the variables we use are of the highest possible level, then as Kinston (1985) warns "Plausible, satisfying and apparently meaningful fantasy may result" (p. 101).

Explanatory methods, even when used with measurable, are totally constrained by the selection of an arbitrary statistical procedure and they play upon possibly incidental patterns in the data. That is, in their attempt to identify the common variance between variables, they are in danger of taking positive advantage of noise and random elements of the data.

The descriptive procedures play mixed roles in science. Graphical methods like scatter plots, histograms and box and whisker plots can be useful tools in many aspects of data analysis. However, the other more "complex" procedures that are used are method-bound and in many cases, arbitrary in their findings leading to theory conflation and confusion.

10.10 Concluding Comment

It is interesting to re-read and reflect on this piece some 30 years after it was written and just under 28 years since my last exchanges on this with Ben. Whilst is has a certain naivete it is possible to see in it threads that have had a profound influence

on how I have approached a career of research and development work. But, in addition, it is with a tinge of disappointment that I note how so many of Ben's observations concerning the unscientific nature of so much so-called statistical modelling remain true today. Moreover, it is unclear whether the rapid recent development in machine learning will lead to an eclipse of this confirmatory approach to science, and to a domination of exploratory methods, especially given the breathtaking expansion of what we now consider to constitute "data."

References

Chalmers, A. (1976). *What is this thing called science?* Brisbane: Queensland University Press.

Eisner, E. W. (1979). *The educational imagination: On the design and evaluation of educational programs*. New York: Macmillan Publishing.

Fowler, W. S. (1962). *The development of scientific method*. New York: Pergamon Press.

Harman, H. (1976). *Modern factor analysis* (3rd Edition Revised). Chicago: University of Chicago Press.

Harré, R. (1976). The constructive role of models. In L. Collins (Ed.), *The use of models in the social sciences* (pp. 16–43). Oxford: Tavistock Press.

Hedges, L. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*, 443–455.

Kinston, W. (1985). Measurement and the structure of scientific analysis. *Systems Research and Behavioral Science, 2*, 95–104.

Phillips, D. C. (1987). *Philosophy, science, and social inquiry*. Oxford: Pergamon Press.

Polkinghorne, D. (1983). *Methodology for the human sciences*. Albany: State University of New York Press.

Valentine, E. R. (1982). *Conceptual issues in psychology*. London: Allen and Unwin.

Chapter 11
Ben Wright: Provocative, Persistent, and Passionate

Trevor Bond

"If I have seen a little further it is by standing on the shoulders of Giants."

Isaac Newton to Robert HookeFeb 5, 1676

11.1 Communicating Invariance

The title of this small tribute to the influence of Ben Wright comes from the dedication page in the second edition of Bond and Fox (2007). I had asked around to entice a few Rasch colleagues to contribute some (suitable, alliterative) suggestions for possible inclusion; most brought a smile to my face, quite a few were not publishable. For many, Ben is the Rasch hero–for others, the Rasch villain. Of course, we much prefer our heroes to be a little more perfect than we are ourselves.

One of Ben's greatest attributes, to my mind, has been his undoubted ability as teacher and communicator. He had a way of elucidating the key Rasch ideas so that they confronted the everyday (mis-)understandings of his audience. Ben's measuring rule was never far from his hand as he pointed out the properties required of scientific measurement, and the inadequacies of what was on offer with true score theory (or IRT). "A rubbery bit here," he would say, mockingly. Or he would ask

T. Bond (✉)
James Cook University, Townsville, QLD, Australia
e-mail: trevor.bond@jcu.edu.au

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_11

about the "missing bit" between the ends of the ruler, or point out their misalignment.

But, most telling of all were his efforts in the face of a confident public critic: Having very easily elicited from that critic willing endorsement of particular basic measurement properties, he would then publicly challenge the naysayer to explain how the more preferred favoured analytical method actually instantiated those principles in practice. Well, of course, as we would expect, it did not. Then Ben would follow with his crystal clear exemplar of how Rasch measurement did.

The most obvious of these revolve around the core concept of measurement *invariance*. Well, of course, the challenger would admit, measures should (must?) be invariant. O.K. So, try comparing the item difficulty calibrations estimated using the more able half of the sample with those derived from the less able half. Yours aren't the same? Shame, that. Ours are. [After all, invariance *is* a fundamental measurement property, and we do work hard to make sure our Rasch measures are invariant.] Now, of course, this revelation would be hard enough to take in the privacy of one's own office, but a public demonstration is a little bit too hard to take. Ben had a disarming way of finding the Achilles' heel of an argument—amusing and salutary to watch, but devastating if you, personally, provided the object lesson.

Those of Ben's qualities listed in the title: provocative, persistent, and passionate (along with others) made him the proselytizer that Rasch measurement apparently needed. While many in our group can list some of those terminally offended by Ben's manner, could Rasch measurement be where it is now, without his *passion*? Many will know that for a long time the work of Ben as well as that of his students was routinely rejected by editors and reviewers of a number of key journals. Someone without his *persistence* would have yielded; but, then, someone less *provocative* would not have upset so many in the first place.

But would so many have even heard the Rasch message without him? There are those among us who claim their own sense of injury directly at Ben's hand; even a few who would claim Ben's mantle as his own. Ben, of course, has been a very human hero. Generous to a fault with praise and support, but stinting in recognizing the benefits of work that stepped outside his tightly proscribed definition of the Rasch measurement bounds. Brilliant performer when centre-stage, but often unwilling to sit back without interjection or running commentary during the presentations of others.

Many Rasch colleagues will remember that for a long while I harboured rather important misgivings about the fit statistics used in Rasch software, such as *Quest* and *Winsteps*. After all, everyone in my field knew that you couldn't make good quantitative indicators of Piagetian cognitive development. American developmentalists, in particular, had made the empirical disproof of Piagetian theory using factor analysis almost an art form. So my presentations of Rasch-based research into cognitive development attracted plenty of attention at Jean Piaget Society meetings in the US. Indeed, it was at JPS meeting where the contact with publishers, Erlbaum was initiated. But, I wasn't looking too closely at whether Rasch fit statistics really worked or not; I was happy enough to bask in a few fleeting moments of sunshine.

11.2 The Value of Theory

Eventually, a prominent US Piagetian, Terrance Brown, Ben Wright and I met at Ben's office in Judd Hall to discuss these data, the Rasch results, the possible interpretations and what sort of impact such results could have for Piagetian theory more broadly. Terry had previously met Ben, professionally, when he worked at the U Chicago medical centre. He also knew from my JPS presentations, that each set of Rasch analysis results that I had presented, was really close to a first attempt at scale construction using those Piagetian based tasks and tests. I was quite relieved when Ben expressed surprise that my students and I had achieved these results straight out of the box; he indicated quite clearly that such first-up results were the exception, rather than the rule, in his experience–and commented that Rasch fit statistics did make life difficult for other researchers.

As we went on to chat about our 'test development' procedures, Ben just shook his head and smiled at the straight-forward approach I adopted with my students: pick an empirical chapter from a suitable Piaget text; develop a coding matrix for children's performances based on Piaget's own exemplars; interview a bunch of suitable children; code the scripts and apply the Partial Credit Rasch model. QED. Ben recognised immediately the advantage we had: a substantive theoretical base of grandiose proportions. Piaget's *oeuvre* consisted of 53 books, and 523 published papers. Is there another similar monument of theory-building empirical research anywhere else in the human sciences?

The upshot of our discussion was to be a series of Rasch workshops attached to the annual JPS meetings. In spite of, or perhaps because of it, Ben insisted that he deliver the first workshop in my absence from the next annual meeting. A number of Piagetian colleagues had the scales fall from their eyes that day: Piaget's substantive theory of human development meets Rasch's theory of measurement for the human sciences. But the follow-up workshops planned for the series were scuttled, so great was the offence taken by a very eminent professor at Ben's apparent dismissal of his very basic, but persistent queries about Ben's robust championing of the Rasch model. Ben had answered repeatedly those same questions from battalions of beginners with indefatigable good humour and patience. The same beginner questions from sages who should know better often provoked his ire.

11.3 Ben's Living Legacy

The Rasch measurement community is a very broad church these days. While most of the prominent practitioners owe their eventual success to that same trait of persistence (it really has been an uphill battle); not all are quite so passionate, and many have determined to be far less provocative than Ben was. Some are careful not even to use the offensive R-word in their papers and presentations, and opt, instead, to couch their work in terms of IRT models. No doubt some of these colleagues will

turn out to be the Trojan horses of the ‘measurement’ community. Nevertheless, the Freudian in me smiles at the stunning successes of those of Ben’s former students whose separation from their mentor was apparently so unpleasant or traumatic. Ben’s other interest in Jungian theory and the functioning of the psyche was always lurking just under the surface.

From Ben we have a legacy of a commitment to a theory of how scientific measurement should function as part of the human sciences; the crucial role of substantive theory in specifying the components of the latent trait underlying scale construction, and the indispensable function of clear, unambiguous communication: A fitting tribute to an all too human hero.

As Rasch measurement researchers, we should take to heart the advice of two of Piaget’s closest *collaborateurs*:

> “If you want to get ahead, get a theory.”
> (Karmiloff-Smith & Inhelder, 1975).

References

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.

Karmiloff-Smith, A., & Inhelder, B. (1975). If you want to get ahead, get a theory. *Cognition, 3*(3), 195–212.

Chapter 12
Benjamin D. Wright: A Higher Standard

Gregory Ethan Stone

Abstract In 1992, I approached Ben Wright about the vexing problem facing the testing organization with whom I worked. The establishment of effective criterion-referenced standards that could be used without inevitable adjustment seemed out of reach when employing most traditional models. Soon Ben would become my professor, mentor and friend. With his irrepressible energy, he quickly produced a wealth of published and, more importantly, unpublished thought on the matter. During the next 4 years I was fortunate enough to work with Ben in the development of what is now called the Objective Standard Setting model. His vision helped assessment to define a new pathway to equity and meaningful measurement. As he had done for so many others before me including my father, his wisdom and inspiration would help me to find a new and passionate career, and to share that experience with new generations. This paper presents the development of criterion-referenced standard setting and the vital role Benjamin Wright would play in this important pursuit.

12.1 Criterion-Referencing Emerges

The notion of criterion referencing in the field of testing has certainly existed for as long as the tests themselves (Binet, 1905). It is impossible to conceive of a mathematics examination for instance, that does not purport to measure some aspect of mathematics. The goal of achievement testing is surely to assess performance against some degree of mastery, however that mastery might be described, and by design the items on tests refer explicitly to the construct under which they are framed.

Robert Glaser beginning in 1962 and subsequently in 1963 introduced the criterion-referenced test as a specific measurement concept. Glaser sought a reasonable understanding of individual behavior through the use of standardized tests.

G.E. Stone (✉)
University of Toledo, Toledo, OH, USA
e-mail: gregory.stone@utoledo.edu

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_12

However, he considered the test scores from classical assessment to be rather poorly elaborated and vaguely specified. Glaser envisioned a continuum of knowledge along which we might locate a person's ability. In 1963 Glaser described this concept as a "continuum of attainment".

> … a student's score on a criterion-referenced measure [continuum] provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what the individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others. (1963, pp. 520).

From Glaser's description came several key ideas that were crucial to the notion of this criterion and its associated measures. First, underlying the "more-to-less" defined criteria was a real and describable construct. Such a construct could represent any content from fundamental mathematics, English grammar, nursing skills to construction worker proficiency. Whatever the content, the construct should clearly be defined as an unambiguous continuum such that on one end exist individuals who possess little of the trait and on the other end exist individuals who possess great quantities of the trait. Along this continuum are infinite shades of performance rather than discrete points, yet for practical use a discrete point would be necessary.

Second, although never fully elaborated, Glaser expected that measures for each student would be independent of one another. The requirement of independence attempted to steer testing bodies away from the traditional practice of normative referencing, determined to be as unreliable a measure as it was an unfair. Glaser's hope tended towards a more systematic approach that would focus on construct idealization rather than on past performance.

In 1962, Mager added one additional component to criterion referencing that would propel the endeavor to its most popular and, according to Glass (1978), ill-conceived destiny. Referring to a more concrete continuum, he suggested that a minimum level of acceptable performance could be specified for each content area. This "performance standard" he determined could be used in the assessment of educational program achievement. Adding the notion of a "minimal level of performance" required the clear dimension of a discrete cutoff point along the specified yardstick. Gene Glass suggested that this addition served to change the focus of standard setting by replacing a behavioral objective with a discrete performance level criterion on a poorly defined variable. The rigidness and non-linearity of the proposed scale were, at the time, considered irreconcilable. The newly defined Rasch model and the perseverance of Benjamin Wright would soon change standard setting in revolutionary ways.

12.2 Standard Setting and the Development of Linear Constructs

At about the same time Nedelsky (1954) was establishing his intricate classification scheme firmly based in a classical approach to testing, measurement as a discipline and as a science was itself undergoing fundamental change. Early pioneers in the measurement of human behavior had uncovered serious problems associated with the relationship between collected data and analytic methods in the social sciences. One of the first problems uncovered related to the data scale itself. Both Thurstone (1927) and Thorndike (1926) expressed concerns with the scale of the collected responses on intelligence tests. They surmised that a linear scale meeting the specification required by most parametric tests was lacking from human response data and their raw scores. In different ways each tried to rectify the situation. Their early successes at linearizing the scale prompted new questions that were not so easily resolved.

While the new linear scales met the algebraic criteria for parametric tests, they were also quite inseparable from the unique sample of persons and items from whence they were obtained. Loevinger (1947) first effectively expressed the idea that measurements of human behavior must be independent of (not overly influenced by) the instruments used—either in the form of persons encountering items or items measuring persons. Angoff (1971) further suggested that the scale would retain its meaning only so long as the groups of persons involved in the process all resembled the group that initially took the test. Changes in population, he reckoned, would become a serious problem for the interpretation and indeed the meaning of scores. It is ironic that Angoff would contemplate this important issue, yet would continue to pursue a classically based standard setting system, which was neither linear nor free of sample interference.

In 1953 Georg Rasch constructed the first complete and decisive system to address these primary concerns. Rasch saw that the probability of an examinee responding correctly to an item must be dominated by only two observable parts to be useful. On the one hand there must be a parameter of ability that relates uniquely and specifically to a person. This ability must be existentially independent of the particular items that are encountered and must therefore be mathematically independent. On the other hand, there must be some level of difficulty associated with an item that exists irrespective of the particular individuals who might encounter the item. Furthermore, these difficulties and abilities must be independent of the other members of the testing cohort or item bank. The paradigm shift represented a clear move from normative information to independent criterion-based information. A person will possess a level of ability that should not be dependent upon the abilities of the other test takers. Similarly, items possess a quality of difficulty that will exist beyond the particulars of the other items that might surround it on an exam paper.

Rasch's application of a logistic response model to the measurement problems described would become the first complete model to meet the three specifications essential for proper measurement (specific objectivity, sufficiency and additivity).

His foundational work would allow many subsequent researchers to build more intricate structures of analysis on a solid foundation. Unlike purveyors of traditional standard setting approaches who worked from the top down and tried along the way to correct for flawed measurement models, the Rasch approach began from the ground floor with a stable and scientifically defensible base model for handling human response data. Only after such grounded foundations were laid could more elaborate systems be built.

12.3 Breaking New Ground

The power of the objective (Rasch, 1960) system was first used for the determination of standards in 1981 by Benjamin Wright and Martin Grosse at the National Board of Medical Examiners. In their initial report, previously established standards (presumably developed from a normative referenced system) were simply converted onto a logit scale. While it did not make full use of the objective system in the determination of the standard, the study did provide highly supportive evidence for the stability of logit standards. Wright and Grosse reported that the variability of failure rates obtained by using a "fixed standard" and those obtained using a norm-referenced standard were not significantly different. This suggested that a criterion point (logit) situated on a linear scale shared a common meaning with any other criterion point from that same scale. The difference that would then exist between the two points would be one of degree only (more ability versus less ability) and not one of construct dimensionality. This discovery was critical. A criterion-referenced standard, if it was to have any meaning at all, must refer to and not deviate from the meaning of the construct developed regardless of the point on the scale chosen as its representative. Non-linear scales (percentages) used by the wide-variety of traditional methods did not fulfill this requirement. The variability of the passing rates using the non-objective models was considerable and excessive.

Wright and Grosse worked for several years on the matter of standard setting, producing a number of published and unpublished papers between 1978 and 1984. However, Hughes, Schumacker and Wright (1984) were the first to systematically investigate multiple standard setting methods and to include a design that more fully exploited the capacities of the objective measurement model. Four methods (Angoff, 1971; Ebel, 1979; Hughes et al. 1984) were used to set criterion standards and were later compared for efficiency of use. The Hughes, Schumacker and Wright (NBME) model should be considered as the initial elaboration of an objective system. Like Angoff, Hughes asked each judge to distinctly define a minimally competent (borderline) individual. Judges then approached the items by speculating about the number of minimally competent individuals would respond to the items correctly. The answer was presented as a probability. The probability in this instance was established for the entire item, irrespective of the response choices. Unlike Angoff, however, Hughes converted the predictions (reported in percentages) to linear measures (logits) using the fundamental Rasch log odd unit. The Rasch model specifies that

the probability of a correct response to a particular test item (P) is controlled by the examinee's ability (b) and the item's difficulty (d). The difference between examinee ability and item difficulty is equal to the log odds (logits) of a correct response, such that:

$$(b - d) = \log[P / 1 - P] \tag{12.1}$$

Hughes ultimately regressed the observed difficulties of the items (obtained from an actual administration) on the predicted logits (from the conversion). The intercept was selected as the minimal standard.

The Hughes, Schumacker and Wright application of the Rasch model is notable because it highlighted the need for linearity. The stability of the measure and the definition of the construct initiated by Wright and Grosse (1981) was firmly established. Without linearity and without the construction of a stable scale the science of measurement and the meaning derived from such pursuits is corrupted. The three authors of the 1984 study found unequivocally that of the four models evaluated, the objectively based method was the most consistent, and by the use of a regression line it was also least sensitive to aberrant judgments about particular items.

While making extensive use of judgment data (converting percentages to logits and thus constructing an adequate measurement tool) the Hughes approach nevertheless continued to rely on indirect information concerning content. Like Angoff and other traditional models, it relied upon judges who were experts within their respective fields to make predictions of performance for mythical minimally competent individuals. Given the similarity in the use of judges between the Hughes and other traditional models, the similarity of outcome was not completely unexpected. The Hughes system, while not as sensitive as Angoff regarding aberrant judgments for particular items and thus not requiring perfect judge agreement, continued to use the predictions of examinee-item interaction as the basis for establishing the criterion point. Such was as theoretically flawed in an objective system as it was in the traditional models.

Grosse and Wright made the first giant leap in the use of judges in a 1987 study, also conducted at the National Board of Medical Examiners. In their new model, judges were asked to participate in a three-phase process. During the first phase, each judge was asked to select a "personal set of criterion items" from a total test of 240 items. Judges were instructed to consider four rules in selecting their criterion set. The rules were specified as follows:

1. The item is highly relevant to practice;
2. The item tests attitudes, skills or knowledge required frequently that should be maintained at an efficient and effective level of quality by every practitioner;
3. The lowest-ability candidate who is clearly certifiable should know the information tested by the item; and
4. The item has only one correct answer.

During the second phase, each judge used his or her own set of criterion items to establish a minimum passing score (a percentage correct required to pass). Finally,

in the third phase, judges were provided with performance information of the items in their individual criterion sets. The information included a breakdown of the percentage of candidates who selected each option of each item. Using the actual performance information, judges could delete some of their originally selected items from the criterion set and adjust their passing percentage if they so chose. The review of items was not undertaken in a manner akin to classical iteration, which promotes acquiescence to norms and reduces judge expertise. Instead, judges were instructed to more closely examine items for problems of syntax, vagueness of wording, and other non-content-based reasons for problematic performance. It was assumed during this process that any of the judge's decisions about content that had been made during the initial selection were acceptable.

The final criterion standard was determined using a version of the Rasch PROX formula:

$$b = H + X \ln[P / 1 - P] \tag{12.2}$$

where
b = the judge's criterion standard for the entire test (in logits)
H = the average difficulty of the judge's criterion items
X = (1 + w2/2.89) 1/2
w = standard deviation of the judge's criterion item difficulties, and
P = percent correct standard set by the judge.

In the early pilot project, standards were set based upon the most difficult item encountered in the criterion set. For the final version, Grosse and Wright chose the PROX formula above which made use of the entire criterion set. This holistic inclusion would later lead Stone (1996, 2004) to the embrace the concept of mastery within the Objective Standard Setting model.

The Grosse and Wright approach represented a major leap beyond the use of judge predictions of success. Instead it was the selected criterion set of items, their content and presentation that defined the construct and the criterion point. Expert judges were for the first time employed in an activity well-suited to their expertise, rather than engaged in a pursuit of speculative prophecy. The judges would select items based on content that was considered important according to their established content guidelines and observations of the profession. Afterwards, actual performances of those items rather than judge speculations would establish the criterion point.

The shift in the use of judges once and for all set Rasch-based standard setting models apart from traditional approaches and established Benjamin Wright as a cornerstone of the movement. In effect, the modern approaches took seriously concerns expressed by Glass (1978) and Jaeger (1979) who had questioned the viability and usefulness of judge predictions. With the benefits of a more reasonable measurement model, Grosse and Wright began a push toward a generally more understandable, meaningful and practical approach. In their reports, they highlighted that the procedures were more understandable to judge participants, required less in the way of training, and were substantially more cost effective—a continual concern to testing bodies.

Additional refinement saw the process simplified further. In Julian and Wright (1988) and Wright and Grosse (1993), the authors describe a relatively simple way to establish a criterion point that hearkened back to the original practical and philosophical conception of the Rasch model. Their suggestion was that there were two principal questions that could be asked of a standard setting judge. Judges could decide whether or not the content and presentation of an item should or should not be required for a test taker to be considered competent. Alternatively, judges could also evaluate the test takers themselves to decide whether or not they should be considered as competent in the assessed content area. These two issues represent the only information that is really available and observable in a testing situation.

Beyond the further general refinement of the Rasch approach, Julian and Wright succeeded in the advancement of another two basic yet often ignored concepts: mathematical simplicity and conceptual understanding. Their process proceeds without complex algebraic formulas and corrections. It makes full use of the straightforward Rasch approach to define and describe a criterion point. Further, it allowed the user to know within reasonable levels of certainty, how precise the standard would be, by supplying error terms obtainable only through a Rasch system.

The Julian and Wright reformulation also simplified the tasks required of judges. Panel experts are selected for their expertise in the content area measured by the examination. These experts are neither measurement scholars nor fortune-tellers. By asking judges to answer questions related to content rather than prediction of performance the task was made significantly more reasonable. Data that are collected through methods that are understandable are clearly more useful and meaningful than those collected through questions that are vague and unanswerable from the outset. The final realization of Julian and Wright established that criterion-referenced standard setting in the spirit and letter of Glaser's consideration was possible and reasonable.

12.4 Progress Continues

Since the foundational work of Ben Wright, modern standard setting has made advances in many directions. Led by the Objective Standard Setting model (Stone, 1996), the first systematic and completely Rasch-based model to be developed, a family of objective measurement methods has been developed. Today, scholars including Matthew Shultz and Mary Lunz have become leaders in the field by advancing other Rasch-based methods including close associates of bookmarking procedures and exercises useful for practical examinations.

Most recently, Objective Standard Setting for Judge-Mediated Examinations was introduced to simplify and clarify polychotomously scored performance ratings. While each day brings new advancements to modern standard setting, none would have been possible without the vision and determination of one man who realized that the best way to improve the model was to promote and celebrate reasonable human evaluation. He debunked the mysticism associated with tradition and refuted

the idea that by simply using mathematics meaningful outcomes would inevitably follow. He replaced fortune telling with specification and algebraic complexity with human interaction. Ben set higher standards in this important work as he did throughout measurement. We must now ensure that we continue that progress moving forward to realize his vision of meaningful measurement. Thanks for leading the way, Ben.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 506–600). Washington, DC: American Council on Education.

Binet, A. (1905). New methods for the diagnosis of the intellectual levels of subnormals. *L'Année Psychologique, 12*, 191–244.

Ebel, R. L. (1979). *Essentials of educational measurement*. Englewood Cliffs: Prentice Hall.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237–261.

Grosse, M. E., & Wright, B. D. (1987). *Criterion item standard setting*. Philadelphia: National Board of Medical Examiners.

Hughes, F. P., Schumacker, C. F., & Wright, B. D. (1984). *Estimating criterion referenced standards for multiple-choice examinations*. Philadelphia: National Board of Medical Examiners.

Jaeger, R. M. (1979). Measurement consequences of selected standard-setting models. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement*. Washington, DC: National Council on Measurement in Education.

Julian, E., & Wright, B. D. (1988). Using the computerized patient simulation to measure the clinical competence of physicians. *Applied Measurement in Education, 4*, 299–318.

Loevinger, J. A. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 61*(4), i.

Mager, R. (1962). *Preparing instructional objectives*. Palo Alto: Fearon Publishers.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3–19.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Stone, G. (1996). *Objective standard setting*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April.

Stone, G. (2004). Objective standard setting (or truth in advertising). In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement*. Maple Grove: JAM Press.

Thorndike, E. L. (1926). *The measurement of intelligence*. New York: Columbia University Teaching College.

Thurstone, L. L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology, 18*, 505–524.

Wright, B. D., & Grosse, M. (1981). *Part II item bank and standard setting study*. Philadelphia: National Board of Medical Examiners.

Wright, B. D., & Grosse, M. (1993). How to set standards. *Rasch Measurement Transactions, 7*(3), 315–316.

Chapter 13
Ben Wright, Rasch Measurement, and Cognitive Psychology

Ryan P. Bowles, Karen M. Schmidt, Tracy L. Kline, and Kevin J. Grimm

Abstract Ben Wright has influenced cognitive psychology both through his own work and through his training of cognitive psychologists. We provide several examples of our efforts to apply the Rasch measurement techniques Ben taught us to cognitive psychology. We describe results from studies employing fit analysis, differential item functioning analysis, Rasch item design techniques, and item linking. These studies address several aspects of human cognition, including spatial visualization, working memory, vocabulary ability, foreign language learning, and cognitive aging. None of these results would be possible without the Rasch measurement techniques we learned from Ben Wright.

13.1 Introduction

In the later part of his career, Ben Wright endeavored to apply Rasch measurement principles in many new areas of scientific research, including clinical psychology (Chang & Wright, 2001), pediatrics (Campbell, Kolobe, Wright, & Linacre, 2002), rehabilitation medicine (Heinemann, Linacre, Wright, Hamilton, & Granger , 1994), and computer adaptive testing (Lunz, Bergstrom, and Wright, 1992). In addition, he trained many young scientists in the use of Rasch measurement techniques, so that they could apply the techniques to new areas and teach more new scientists. Ben's interest in teaching and challenging young scientists to think clearly about

R.P. Bowles (✉)
Department of Human Development and Family Studies, Michigan State University, East Lansing, MI, USA
e-mail: bowlesr@msu.edu

K.M. Schmidt
Department of Psychology, University of Virginia, Charlottesville, VA, USA

T.L. Kline
RTI International, Research Triangle Park, Durham, NC, USA

K.J. Grimm
Department of Psychology, Arizona State University, Tempe, AZ, USA

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*, Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_13

measurement in their own research domains remained outstanding very late in his career (Wright, 1999). The authors of this chapter are examples of Ben's success in propagating Rasch methodology in different research fields. We have applied the knowledge Ben taught us to inform our understanding of human cognition in several domains, including verbal and spatial ability, working memory, vocabulary ability, foreign language learning, and cognitive aging.

Before we describe how Ben influenced our work, we would be remiss not to mention that, as has been the experience for many researchers in other fields that Ben has influenced, Ben completed research in cognitive psychology that predates ours. Every user of Winsteps is familiar with Ben's work with the Knox Cube Test, a measure of spatial memory (Stone & Wright, 1983; Wright & Stone, 1979). More recently, the designers of the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R; Woodcock and Johnson, 1989) consulted with Ben to develop the only comprehensive test of cognitive abilities based on Rasch measurement principles (McGrew, Werder, & Woodcock, 1991, Chap. 3). In fact, the W scale, in which scores on the WJ-R are reported, is in part named for Ben Wright (R. W. Woodcock, personal communication, April 18, 2003). Thus, Ben has influenced cognitive psychology both directly through Rasch-scaled cognitive tests, and indirectly through his training of cognitive psychologists.

This chapter describes some of our efforts in using Rasch measurement techniques in cognitive psychology. In particular, we describe studies that have used fit analysis, differential item functioning analysis, Rasch item design techniques, and item linking. We first describe two studies that employ fit analysis, one that examines the Block Design task on the Wechsler Adult Intelligence Scale- Revised (WAIS-R; Wechsler, 1981) to understand spatial visualization, and one that considers the synthesis of multiple skills in foreign language learning. We then describe two studies that use differential item functioning analysis, one that considers age differences in proactive interference and one that examines strategy differences in spatial visualization. Next, we describe two studies that address issues in item design for cognitive tests: the Spatial Learning Ability Test, which measures spatial visualization; and the Object Location Memory Revised test, which measures memory for spatial locations. Finally, we describe a study that uses item linking in order to measure vocabulary ability throughout the lifespan. Together, these studies highlight the importance of Ben Wright and fundamental measurement in our understanding of human cognition.

13.2 Fit Analysis

Good measurement occurs only when the data fit a measurement model, not when the model fits the data, as Ben often emphasized (Wright, 1977, 1994). When the data do not fit the Rasch model, the pattern of misfit can often be very informative (Linacre & Wright, 1994). Fit information can help identify patterns of misfit that have meaningful interpretations for understanding the psychological processes involved in responding to an item. We provide two examples of the use of fit

information in cognitive psychology. In the first, we describe how fit statistics were used to help identify why the Block Design task is highly diagnostic of deficits in executive functioning. In the second example, fit statistics were used to help understand how foreign languages are learned.

13.2.1 Block Design

The Block Design task on the WAIS-R (Wechsler, 1981) consists of ten visual patterns that can be created from a set of colored blocks. The test-taker is given a set of blocks and must recreate the pattern within a time limit. Figure 13.1 provides an example of an item similar to those on the Block Design task. A successful completion is scored as 4 points, while failure is scored 0. On the first two items, the examinee is shown how to put the blocks together, and then has to replicate it. If unsuccessful, a second trial with a second demonstration is given, with a score of 2 given for success. On the last four items, 1, 2, or 3 bonus points are awarded for increasingly fast successful completions.

Block Design is very sensitive to cognitive deficits, as a result of both central nervous system dysfunction (Lezak, 1995) and aging (Kaufman, 1990; Troyer, Cullum, Smernoff, & Kozora, 1994). The reasons why Block Design is highly diagnostic of many types of cognitive deficits are not known, although many hypotheses have been tested and supported (Joy, Fein, Kaplan, & Freedman, 2001; Salthouse, 1987; Storandt, 1977; Troyer et al., 1994; Wilde, Boake, & Scherer, 2000). Some researchers have suggested that Block Design is psychologically complex, and taps many different processes, which are required to differing degrees across items (Kaplan, 1988; Kaplan, Fein, Morris, & Delis, 1991). This hypothesis implies that the Block Design test is multidimensional, and should not fit the Rasch model, perhaps in predictable ways.

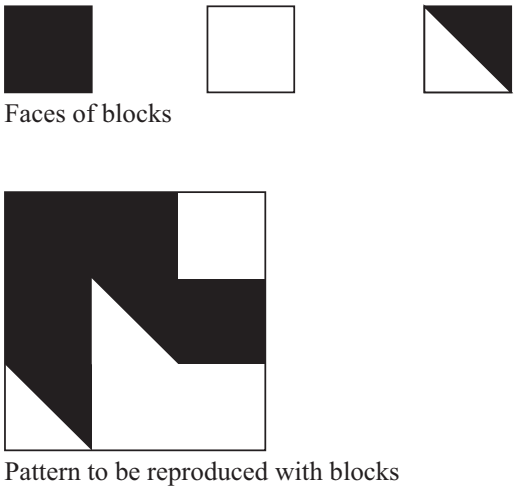


Fig. 13.1 Example of Block Design blocks and pattern

Table 13.1 Item fit for Block Design with time bonuses

	Infit		Outfit	
Item	MNSQ	ZSTD	MNSQ	ZSTD
Block Design 1	0.38	−8.7	0.37	−7.5
Block Design 2	0.37	−9.0	0.33	−8.2
Block Design 3	0.45	−7.3	0.41	−6.8
Block Design 4	0.35	−9.2	0.30	−8.8
Block Design 5	0.40	−8.4	0.34	−8.1
Block Design 6	0.43	−7.9	0.38	−7.3
Block Design 7	1.32	3.1	1.36	2.9
Block Design 8	1.88	7.7	1.97	6.8
Block Design 9	1.99	8.8	2.26	8.2
Block Design 10	1.95	8.3	2.27	7.8

Table 13.2 Item fit for Block Design without time bonuses

	Infit		Outfit	
Item	MNSQ	ZSTD	MNSQ	ZSTD
Block Design 1	0.93	−0.1	0.55	−0.2
Block Design 2	0.76	−1.0	6.65	2.1
Block Design 3	1.43	0.8	9.90	3.4
Block Design 4	0.49	−1.2	0.04	−0.8
Block Design 5	1.18	0.5	1.40	0.2
Block Design 6	0.99	0.0	3.90	1.7
Block Design 7	0.87	−0.7	4.35	2.4
Block Design 8	0.82	−1.3	0.72	−0.7
Block Design 9	0.75	−2.3	0.45	−1.2
Block Design 10	1.10	1.0	0.76	−0.3

As part of the National Growth and Change Study, Bowles and McArdle (2000) analyzed Block Design data from 149 people using the Rating Scale Model (RSM; Andrich, 1978) in Winsteps (Linacre & Wright, 2001). As can be seen in Table 13.1, the final four items show a different pattern of misfit than the first 6 items. Because the last four items involve the time bonuses, these results suggest that speed in solving Block Design items does not reflect the same type of cognitive processes as those involved in simply recreating the design. In a second analysis, the time bonuses were removed from the scoring. Item statistics are presented in Table 13.2. The Block Design items still misfit the RSM, and there is no apparent pattern in the misfit. Furthermore, the pattern of misfit did not match any of the sets of items that previous research has identified as involving different cognitive processes (items 1, 4, and 6, Kaplan et al., 1991; items 1, 5, 6 and 8, Joy et al., 2001; items 2, 7, and 9, Joy et al., 2001). These results suggest that, although Block Design is highly predictive of cognitive deficits in general, it does not measure a single coherent dimension and is not likely to be useful for identifying specific types of deficits.

13.2.2 Foreign Language Learning

MultiCAT (Ohio State University Foreign Language Center, 2002) is a series of two Rasch-based adaptive tests in three languages designed to measure aspects of second language proficiency for college placement and exit proficiency. The Reading test contains items that consist of a reading passage and a single multiple choice comprehension question. The Reading test has been through extensive calibration testing, and excellent fit to the Rasch model has been established. However, a small number of items have been identified as misfitting. Ongoing research examining the types of misfitting Reading items indicates that items involving both vocabulary knowledge and grammar knowledge tend to misfit, while items involving one or the other do not. This result suggests that the synthesis of grammar and vocabulary is a separate dimension of foreign language learning than grammar or vocabulary alone. Although this research is still in its preliminary stage, the current results point to a direction for research into the way people learn foreign languages.

13.3 Differential Item Functioning

Differential item functioning was identified as an important issue in measurement by Ben many years ago (Wright, Mead, & Draba, 1976). As Wright et al. pointed out, on a test with good measurement properties, the meaning of ability and difficulty "can only be the consequence of the person's and the item's position on the trait and so they must hold regardless of the race, sex, etc. of the person measured." However, when there are group differences in item difficulty, the differences can be informative. We provide two examples of how examining differential item functioning can inform our understanding of human cognition. In the first, we tested a theory about the aging of working memory that yields predictions about age group differences in item difficulty. In the second, we describe how exploring differential item functioning on a test of spatial visualization can lead to important insights into individual differences in item solution strategy.

13.3.1 Proactive Interference and the Aging of Working Memory

Working memory is a system for the simultaneous storage and manipulation of information. The amount of information that can be simultaneously stored and processed, known as working memory span, is limited, and declines with age (Verhaeghen & Salthouse, 1997). Increased susceptibility to proactive interference (PI) has been suggested as one cause of the age-related decline in working memory span (Hasher & Zacks, 1988; May, Hasher, & Kane, 1999). PI is a reduction in the ability to perform a cognitive task because of interference from prior performance of the same or a related task. PI may build up over the course of a working memory span task, with

Fig. 13.2 Age-group differences in working memory span item difficulties. From Bowles and Salthouse (2003)

the first trial having no effect of proactive interference, the second trial having PI from the first trial, the third trial having PI from the first and second trials, etc. If older adults are more susceptible to the effects of PI than younger adults, then later trials should be relatively more difficult for older adults than for younger adults.

We examined this prediction by examining differential item functioning across age groups with the Rasch model (Bowles & Salthouse, 2003). Two working memory span tasks were given to 698 persons, who were divided into three age groups, young (age < 40), middle (between 40 and 59 inclusive) and old (age ≥ 60). For both tests, results supported the prediction that later presented items would be relatively harder for older adults than for younger adults (see Fig. 13.2 for results from one of the tasks). Furthermore, the variance shared by age and WM span was reduced by approximately 50% after accounting for differential susceptibility to PI, indicating that an age-related increase in susceptibility to PI may account for as much as half of the age-related decline in WM span.

13.3.2 Spatial Visualization and Individual differences in Item Solution Strategy

Spatial visualization is the ability to manipulate visual images mentally (see Carroll, 1993, for a summary). Schmidt McCollam (1998) analyzed responses of 211 Air Force recruits to spatial visualization items involving the mental folding of an

unfolded cube, to understand differential strategy application in mental rotation and folding cognitive processes. The Mixed Rasch Model (Rost, 1990), which explores differential item functioning when group membership is unknown, was applied to the data. Two latent classes emerged; further investigation of the patterns of responses indicated that one class (high transform group) excelled on items requiring complex transformation, the other (low transform group) excelled on items requiring relatively simple transformation. When the analysis was extended to explore group differences in scores on subtests of the Armed Services Vocational Aptitude Battery (ASVAB; Bayroff & Fuchs, 1968), we found that the high transform group's scores on electrical information, auto and shop information, and mechanical reasoning were significantly greater than those for the low transform group. No other ASVAB scores were different for the two classes. Hence, the exploratory differential item functioning analyses using the Mixed Rasch Model revealed potential spatial visualization solution strategy differences for separate groups of persons.

13.4 Designing Items

The importance of item design in operationalizing and understanding a construct was often emphasized by Ben. In fact, Ben described *Rating Scale Analysis* as a book "about how to construct variables and how to use them for measuring" (Wright & Masters, 1982, p. 1). Items must be designed to measure a single one-dimensional construct. Beyond the requirement of unidimensionality, items can be designed so that, by manipulating specific aspects of the items, hypotheses about the construct can be assessed. We provide two examples of the use of item design to understand human cognition. In the first, manipulations were introduced to induce performance change in a spatial visualization test. In the second example, manipulations were introduced to items measuring the ability to remember an object's location and identity, to test several hypotheses about how people use spatial memory.

13.4.1 Spatial Learning Ability Test

The items on the Spatial Learning Ability Test (SLAT; Embretson, 1991) require the examinee to mentally fold an unfolded cube and match it to a representation of the folded cube. The items on the SLAT are designed with two fully crossed factors, each with three levels: Degrees Rotation in the plane (0, 90, 180 degrees), and Surfaces Carried in depth (1, 2, 3 surfaces), yielding nine item complexity types (see Fig. 13.3 for an example of a 0-degree, 1-surface item). According to spatial visualization processing theory, increases in degrees rotation (Shepard & Metzler, 1971) and number of surfaces carried (Shepard & Feng, 1972) result in greater solution complexity. Hence, a 180-degree, 3-surface SLAT item should be much more difficult than a 0-degree, 1-surface item.

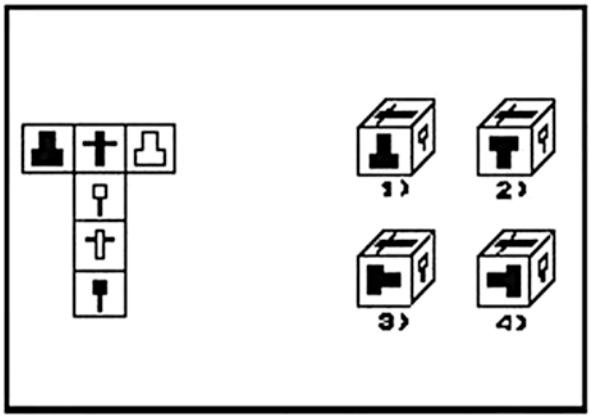


Fig. 13.3 Example of a 0-degree, 1-surface SLAT item

These hypotheses were tested by fitting the data to the Linear Logistic Test Model (LLTM; Fischer, 1973), which combines a Rasch model with a linear predictor of item difficulty. The linear predictor consisted of linear and quadratic effects of both Surface Carried and Degrees Rotation. Results indicated that linear Surfaces Carried (0.71) contributes the most weight to the predicted item difficulty, followed by linear Degrees Rotation (0.35). The quadratic effects were small (0.10 and 0.03, respectively). Hence, the more difficult processing factor for SLAT is Surfaces Carried, and, the effects of the two factors on cognitive processing demands are linearly related to SLAT performance, but not quadratically.

Further studies involving the SLAT used an extension of the Rasch model (Embretson, 1984) to understand lifespan differences in cognitive processes (i.e., general executive function and working memory capacity) underlying test performance (Embretson & Schmidt McCollam, 2000b; McCollam, 1997). One hundred seventy-eight older and younger adults were measured on the SLAT across three testing blocks, with cognitive strategy training given between each administration. Results indicated that general executive function (R^2 = 52%) was more important than working memory capacity (R^2 = 14%) for understanding lifespan differences in spatial processing on the SLAT.

Results from another extension of the Rasch model (Embretson, 1991) showed that age differences in SLAT performance change as a result of the cognitive strategy training were quite different and more meaningful than traditional raw gain scores (Embretson & Schmidt McCollam, 2000a). Specifically, older adults showed more positive change than younger adults after two strategy training periods, while traditional raw gain scores showed the opposite effect. The source of these differences lies in the fact that Rasch-based measurement is on an interval scale, and raw gain scores are on an ordinal scale (Perline, Wright, & Wainer, 1979).

13.4.2 Object Location Memory—Revised

The Object Location Memory—Revised (OLM-R; Kline & Schmidt, 2005) is a task designed to measure complex spatial memory. The task, based on previous work by Silverman and Eals (1992), measures a person's attention to both the appearance and location of a particular image within an array (see Fig. 13.4). In the OLM-R test, participants study an array of items for 30 s, then are presented with a distractor task. After 30 s of the distractor task, participants are presented with a modified version of the study array. Participants have 1 min to indicate which items have been manipulated within the array, either by movement in Cartesian space or replacement with a new image. The moved objects assess the ability to identify changes in object location assignment, and replaced objects assessed the ability to identify changes in image appearance without location cues.

The OLM-R was administered to 114 persons, and the data was analyzed using Winsteps (Linacre and Wright, 2001) and Facets (Linacre, 1989). Winsteps was used to investigate measurement properties of the OLM-R, while Facets was used to investigate the effectiveness of the predetermined item complexity factors. Results from the Facets analysis are presented in Fig. 13.5. It was found that the OLM-R test

Fig. 13.4 Example of a 20-item, spatial distractor, OLM-R array

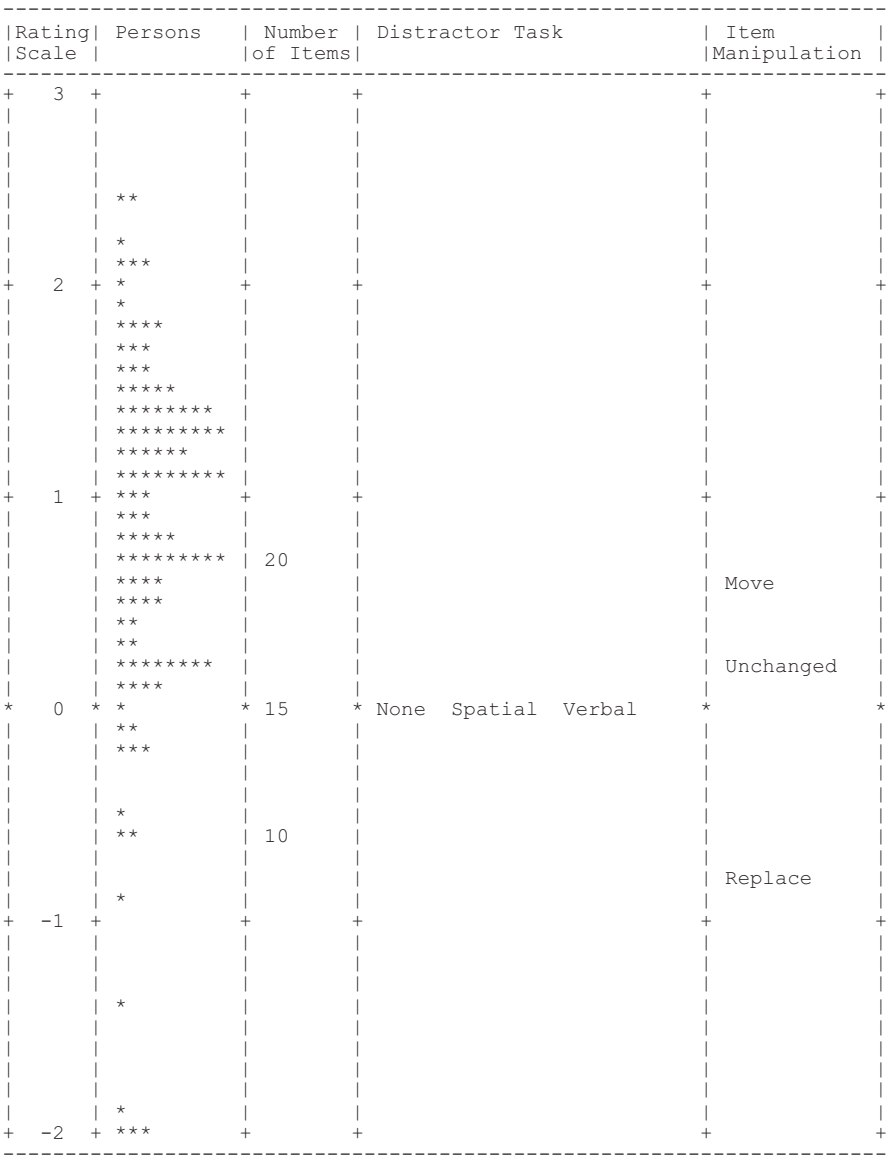


Fig. 13.5 FACETS Wright plot depiction of person ability and factor difficulty

had excellent measurement properties, as the data fit the Rasch model well (Kline & Schmidt, 2005). Furthermore, arrays with a greater number of items were more difficult, suggesting that memory load is an integral component influencing performance on the OLM-R task. The type of distractor task had no effect on performance, indicating that spatial memory may require different processes than typical spatial distractor tasks. Another possibility is that the distraction duration of 30 s was insuf-

ficient to influence performance. Additionally, results on object manipulation indicated that participants encoded item identification (replaced items) with more difficulty than item location (moved items). This suggests that spatial memory may be space-centered instead of item-centered; that is, participants look at the array as a whole, instead of focusing on individual images.

13.5 Linking Tests

As Ben noted, "the quantitative study of development depends on the ability to make measurements over a wider range of difficulty values than can be covered with a single test" (Wright, 1977, p. 108). Assessing change in a cognitive ability requires that the ability be measured on a common scale at all measurement occasions. Otherwise, "change the items, and you have a new yardstick" (Wright and Stone, 1979, p. xi). We provide an example of linking tests with the Rasch model. Several vocabulary tests were linked to yield a single yardstick for vocabulary ability, so that we can understand how vocabulary ability changes with age.

13.5.1 Vocabulary Ability Across the Lifespan

The goal of a recent study (McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009) was to model the lifespan development of vocabulary ability using growth curve analysis with the available cognitive data from the Bradway-McArdle Longitudinal and the Berkeley Growth Study. These longitudinal studies began in the late 1920s and early 1930s and have continued to the present day with the most recent data collections occurring in 2000. Measures of cognitive ability were administered up to nine times during the 70 years of these studies. Over the course of these studies, the researchers have consistently administered different cognitive batteries because of age-appropriateness and revised test batteries. As a result, nine different vocabulary tests were used to measure the vocabulary ability of the participants. Before any conclusions about longitudinal changes in vocabulary ability can be made, a single measurement scale is necessary so that changes in ability can be separated from changes in the tests.

In order to model lifespan changes in vocabulary ability, all of these tests that measure vocabulary ability were put on a common scale using the Rasch model with common person and item equating. The results of the item analysis demonstrated that the vocabulary data fit the measurement model well, lending support for unidimensionality and for modeling the estimated person abilities by the participant's age at testing. Figure 13.6 is the plot of estimated person ability against testing age. In this plot, each line represents an individual, which allows for the visualization of the developmental trajectory of vocabulary from age 4–75.

A dual-exponential growth model (McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002) best represented the lifespan development of vocabulary ability.

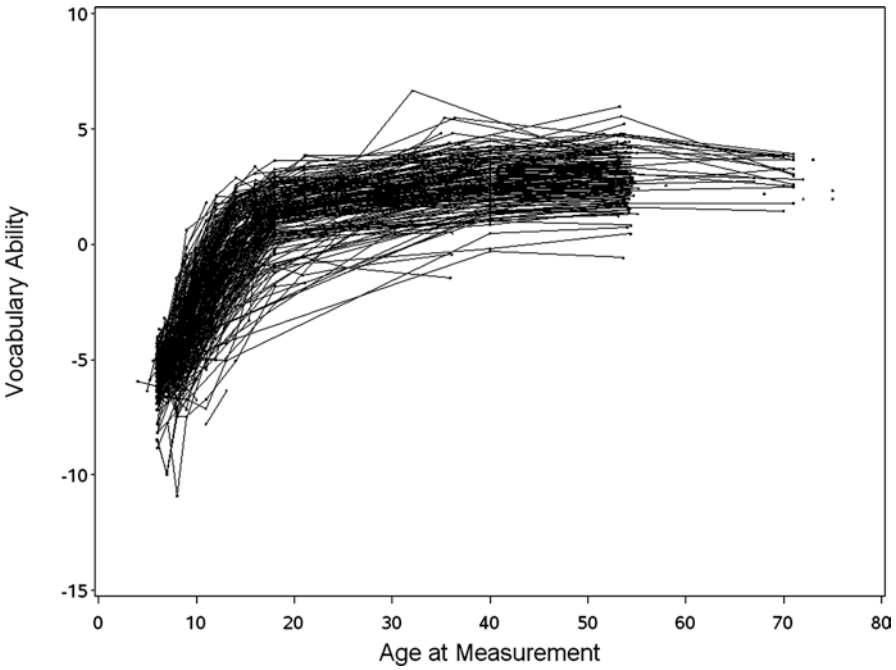


Fig. 13.6 Longitudinal plot of vocabulary ability measure by age at time of testing

In the dual exponential model there is a growth rate and a decline rate. The growth rate accounts for the shape of development during childhood and adolescence, while the decline rate models the changes occurring through adulthood. The growth rate in the dual exponential model was 0.14, while the decline rate was 0.0001, indicating strong but decelerating growth in vocabulary ability during childhood before the changes in ability level off in the mid-thirties and slowly decline into older adulthood. These results confirm previous research on vocabulary ability (McArdle & Nesselroade, 2003; McArdle, Hamagami, Meredith, & Bradway, 2000), which found that vocabulary grows into the early thirties before leveling off, with a very small decline in late adulthood.

13.6 Conclusion

We have developed a better understanding of the way humans think by using Rasch measurement techniques. Good measurement is necessary before valid conclusions about human cognition can be reached, as illustrated by our examples about the Spatial Location Ability Test, the Object Location Memory Revised test, and vocabulary ability across the lifespan. Violations of good measurement can provide further information, as illustrated by our examples about Block Design, foreign

language learning, the aging of working memory, and item solution strategies in spatial visualization. None of this understanding would be possible without Rasch measurement techniques and the person who taught us to use them, Ben Wright.

References

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

Bayroff, A. G., & Fuchs, E. F. (1968). The armed services vocational aptitude battery. *Proceedings of the Annual Convention of the American Psychological Association, 3*, 635–636.

Bowles, R. P. & McArdle, J. J. (2000). An Item Response Theory (IRT) analysis of WAIS and WJ-R sub-scales. In McArdle, J. J (Ed.), *A summary of recent results from the National Growth and Change Study (NGCS)*. Department of Psychology, University of Virginia, Appendix to NIH Grant AG7467, National Institute on Aging.

Bowles, R. P., & Salthouse, T. A. (2003). Assessing the age-related effects of proactive interference on working memory span tasks using the Rasch model. *Psychology and Aging, 18*, 608–615.

Campbell, S. K., Kolobe, T. H. A., Wright, B. D., & Linacre, J. M. (2002). Validity of the test of infant motor performance for prediction of 6-, 9- and 12-month scores on the Alberta Infant Motor Scale. *Developmental Medicine and Child Neurology, 44*, 263–272.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Chang, C.-H., & Wright, B. D. (2001). Detecting unexpected variables in the MMPI-2 social introversion scale. *Journal of Applied Measurement, 2*, 227–240.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49*, 175–186.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 52*, 495–516.

Embretson, S. E., & Schmidt McCollam, K. M. (2000a). A multicomponent Rasch model for measuring covert processes: Application to lifespan ability changes. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. Vol. 5, pp. 203–218). Norwood: Ablex.

Embretson, S. E., & Schmidt McCollam, K. M. (2000b). Psychometric approaches to understanding and measuring intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 423–444). New York: Cambridge University Press.

Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Hasher, L., & Zacks, R. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. Vol. 22, pp. 193–226). New York: Academic Press.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., & Granger, C. (1994). Measurement characteristics of the functional independence measure. *Topics in Stroke Rehabilitation, 1*, 1–15.

Joy, S., Fein, D., Kaplan, E., & Freedman, M. (2001). Quantifying qualitative features of block design performance among healthy older adults. *Archives of Clinical Neuropsychology, 16*, 157–170.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function: Research, measurement, and practice* (pp. 129–231). Washington, DC: American Psychological Association.

Kaplan, E., Fein, D., Morris, D., & Delis, D. (1991). *The WAIS-R as a neuropsychological instrument*. San Antonio: Psychological Corporation.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence*. Boston: Allyn and Bacon.
Kline, T. L., & Schmidt, K. M. (2005). Rasch analysis examining processing mechanisms of the object location memory test revised. *Journal of Applied Measurement, 6*, 382–395.
Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). Oxford: Oxford University Press.
Linacre, J. M. (1989). *Facets [Computer software]*. Chicago: MESA Press.
Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions, 8*, 350.
Linacre, J. M., & Wright, B. D. (2001). *Winsteps (Version 3.02) [Computer software]*. Chicago: MESA Press.
Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16*, 33–40.
May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory and Cognition, 27*, 759–767.
McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology, 38*, 115–142.
McArdle, J., Grimm, K., Hamagami, F., Bowles, R., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14*(2), 126–149.
McArdle, J. J., Hamagami, F., Meredith, W., & Bradway, K. P. (2000). Modeling the dynamic hypotheses of Gf-Gc theory using longitudinal life-span data. *Learning and Individual Differences, 12*, 53–79.
McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary research. In J. Schinka & W. Velicer (Eds.), *Comprehensive handbook of psychology*, *Research methods in psychology* (Vol. Vol II, pp. 447–480). New York: Pergamon.
McCollam, K. M. (1997). The modifiability of age differences in spatial visualization (Doctoral dissertation, University of Kansas, 1997). *Dissertation Abstracts International,* 59*,* 1409.
McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). *WJ-R technical manual*. Allen: DLM.
Ohio State University Foreign Language Center. (2002). *MultiCAT*. Columbus: Ohio State University.
Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*, 237–256.
Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.
Salthouse, T. A. (1987). Sources of age-related individual differences in block design tests. *Intelligence, 11*, 245–262.
Schmidt McCollam, K. M. (1998). Latent trait and latent class models. In G. M. Marcoulides (Ed.), *Modern methods for business research* (pp. 23–46). Mahwah: Erlbaum.
Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive Psychology, 3*, 338–243.
Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701–703.
Silverman, I., & Eals, M. (1992). Sex differences in spatial abilities: Evolutionary theory and data. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 487–503). New York: Oxford University Press.
Stone, M. H., & Wright, B. D. (1983). Measuring attending behavior and short-term memory with Knox's cube test. *Educational and Psychological Measurement, 43*, 803–814.
Storandt, M. (1977). Age, ability level, and method of administering and scoring the WAIS. *Journal of Gerontology, 32*, 175–178.
Troyer, A. K., Cullum, C. M., Smernoff, E. N., & Kozora, E. (1994). Age effects on block design: Qualitative performance features and extended-time effects. *Neuropsychology, 8*, 95–99.

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin, 122*, 231–249.

Wechsler, D. (1981). *Wechsler adult intelligence scale- revised*. New York: Psychological Corporation.

Wilde, M. C., Boake, C., & Sherer, M. (2000). Wechsler adult intelligence scale- revised block design broken configuration errors in nonpenetrating traumatic brain injury. *Applied Neuropsychology, 7*, 208–214.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson psycho-educational battery-revised*. Allen, TX: DLM.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

Wright, B. D. (1994). Data analysis and fit. *Rasch Measurement Transactions, 7*, 324.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah: Erlbaum.

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

Wright, B. D., Mead, R., & Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model (MESA Research Memorandum No. 22)*. Chicago: MESA Psychometric Laboratory.

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

Chapter 14
Provoking Professional Identity Development: The Legacy of Benjamin Drake Wright

William P. Fisher, Jr.

Abstract Ben Wright's background in physics and Freudian psychoanalysis, working alongside wide-ranging, deep thinkers attuned to cross-disciplinary matters, like Charles Townes, Bruno Bettelheim, and Ben Bloom, set the stage for creative engagements with educational problems that still resonate with researchers and practitioners, globally. In Rasch's models for measurement, Wright found a means not only for developing his own professional identity and writing his own life story but for also providing others with the means and media for their own imaginative variations on an invariant.

14.1 Equating Life with Stories

Plato speaks of public events—dramatic theatre, the Olympics, or political debate—as effective in two ways. First, they must address each of us as individuals, giving expression to our private joys and sufferings in a forum shared by all, even if the exact details of the story told do not in fact apply to anyone. Second, they must provide an effective model for meeting the challenges faced by the society, recasting the past to make sense of the present and to see a way forward in the future. In contrast with the ancient Roman concept of the spectator, the ancient Greeks conceived the observer's role as a participant who both shapes and is shaped by the unfolding event.

Ben Wright similarly often spoke of tests and surveys as conversations in which all participants contributed to the telling of a common story and could see where they stood relative to everyone else, qualitatively and quantitatively. How did he arrive at this conception? Bouchard's chapter in this volume recounts how Ben "went in search of life" as a young man (Wright, 1988b, p. 25), and how he had been oriented toward an experimental approach to life as a boy. Ben was deliberate

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-3-319-67304-2_16

W.P. Fisher, Jr. (✉)
Graduate School of Education, University of California, Berkeley, Berkeley, CA, USA
e-mail: wfisher@berkeley.edu

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_14

in his experimental approach to discovering and becoming the person he wanted to be. He had the formal background in psychology he needed to find his way. His training in psychoanalysis in the 1950s familiarized him with Freud's use of mythology in the characterization of psychological processes (the Oedipus conflict) and pathologies (Narcissism, for instance). Wright (1959a, pp. 368–371) reviews Freud's work on identity development in the context of discerning a typology of teacher personalities. Ben's self-aware engagement with these issues of identity development led to some striking results over the course of his career.

Ben's search for life was, then, also a quest for narrative, to adopt Ricoeur's (1991a) title. As pointed out by Ricoeur (1991b, p. 194; also see Somers, 1994), "We equate life with the stories we can tell about it." Ben struggled to find ways to write his life story—and to support others in writing theirs—that would be, as Ricoeur (1991b, p. 196) describes them, "imaginative variations on an invariant." That is, we know ourselves and develop our identities as stable actors in the world indirectly, in Ricoeur's (1991b) terms,

> by the detour of the cultural signs of all sorts which are articulated on the symbolic mediations which always already articulate action and, among them, the narratives of everyday life. Narrative mediation underlines this remarkable characteristic of self-knowledge—that it is self-interpretation (p. 198).

Key to grasp here is the fact that "...the recounted story is always more than the enumeration, in an order that would be merely serial or successive, of the incidents or events that it organizes into an intelligible whole" (Ricoeur, 1991a, p. 21). It will not suffice to simply tally up counts; the whole is inherently greater than the sum of the parts. Intentionally or not, our life choices narrate a history, a story of values, desires, and meanings. One of the most important ways we represent ourselves as who we are is through our choice of profession. As is implied by the word itself, work entails professing certain beliefs, opinions, and understandings. In no career choice is this more the case than in the work of a professor, as Ben was.

The stories we tell about ourselves require a plot. Ricoeur (1991a, p. 21) characterizes emplotment as an "integrating process." He observes that fitting disparate events into an overall narrative "gives a dynamic identity to the story recounted: what is recounted is a particular story, one and complete in itself" (p. 21). This identity is not homogenous or completely uniform, but necessarily incorporates discordances along with concordance.

As Wright came to see, at least implicitly, as events unfolded in his life, variations on an invariant giving a dynamic identity to a story are an apt characterisation of Rasch's (1960, 1977; Andersen, 1977; Andrich, 2010) probabilistic, individual-level approach to statistical sufficiency. Mathematically, statistical sufficiency and invariance are analogous ways of formulating criteria for unidimensional, separable model parameters (Hall, Wijsman, & Ghosh, 1965; Arnold, 1985). Guttman, presenting influential measurement ideas in the 1940s and 1950s, demanded perfect statistical sufficiency and data consistency in his approach to measurement, so that a purely concordant story of each performance can be told. Thurstone and Thorndike, working in the 1920s, similarly achieved some important goals, but remained

limited in the generalizability of their invariants to other samples and tests. Rasch's probabilistic approach, in contrast, focuses attention on unexpected individual responses, bringing them to attention for possible action, at the same time that it sets up a larger basis for generalized comparability over time and across students. Rasch, Thurstone, Thorndike, and Guttman differ, then, in their capacity to tolerate and make use of both discordant and concordant variations on the invariant in the stories told of the measured performance. (See the chapter by Engelhard in this volume for more on these contrasts; also see Andrich, 1978; Engelhard, 1984, 1994, 2012.)

Ben would prove sensitive to these issues, as will be shown. Interestingly, the writer of a story, the one choosing the path in life that simultaneously composes and interprets a narrative arc, is guided by expectations concerning the outcome, and has to readjust those expectations, as much as any reader does. Well-told stories have something to teach us concerning "universal aspects of the human condition," as Aristotle pointed out (Ricoeur, 1991a, p. 21). Given Ben Wright's lifelong immersion in the psychology of teacher identity and the conceptualization of measures as imaginative variations on invariants, we might expect Ben's choices and written records to have a particularly rich story to tell of both his life and the human condition in general.

That, then, is the structure of the task we face. The question is: How do we tell the story of Ben's particulars in a way that speaks to all but remains true to him? Of course, not everyone affected by Ben knows it. And undoubtedly the vast majority of people who will be affected by Ben's work have not yet been touched by it, and have not yet even been born. So, as Ben (Wright, 1997, p. 34) would say, our task is to understand historical data that will never be produced again in the same way to an exact degree of detail, and to do so in a way that enables us to better manage the future.

An advantage I have in trying to start to recount Ben's story in this way is a collection of most of Ben's earliest publications, from the period 1948 to 1968. It is both an exciting opportunity, and a humbling responsibility, to have these papers at hand as resources to draw on. At the very least, this effort will provide a start at formulating a fuller account of Ben's career and contributions.

In his measurement work, Ben made revolutionary innovations in several fundamental areas, such as estimation methods, model design, fit and reliability statistics, software, theory development, and applications in who knows how many fields. Over the course of his academic career, he was intensely interested in students and gave unremittingly of his time and energy to anyone interested in learning what he had to teach.

In addition, over the years Ben was vitally involved in bringing out not only his own work, but that of his teacher, as he was instrumental in the 1980 reprinting of Rasch's 1960 book by the University of Chicago Press. He expanded others' publishing opportunities via the institution of *Rasch Measurement Transactions*, the *Journal of Outcome Measurement*, and other media, such as the series of MESA Psychometric Laboratory Research Memoranda. Ben was also editor of the *School Review*, and on the editorial boards of *The Elementary School Journal* and the *Journal of Educational Measurement* for many years. Finally, Ben was also an

organizer, being a key player in organizing the first ever AERA training presession in Los Angeles in 1969 (which was on Rasch's models for measurement), and in the beginnings of the AERA Rasch Measurement SIG, the Institute for Objective Measurement, the Midwest Objective Measurement Seminars, the Chicago Objective Measurement Table, and the International Objective Measurement Workshops.

The relevant categories in Ben's measurement work would then seem to be methods and theory, teaching, publishing, and professionalization. A common theme across all of these categories is improved access to measurement. Ben introduced parameter estimation algorithms (Wright & Panchapakesan, 1969; Wright & Douglas, 1977; Wright, 1988a) that were faster and more efficient than others then available, a key accomplishment in the technological environment of the 1960s. He was also among the first to write fundamental measurement software that not only worked (Hambleton & Cook, 1977, pp. 76, 88), but was much more informative than cryptic (see Linacre 2017 for the latest version). Ben knew good ideas when he saw them, and adapted reliability coefficients (such as Andrich, 1982) and model fit statistics into his software (Wright & Panchapakesan, 1969; Wright, 1977; Wright & Masters, 1982, pp. 91–92, 105–106, 113). He supported or was directly involved in developing rating scale, partial credit, multifaceted, and other models capable of testing and estimating parameters for virtually any kind of data typically gathered in the human sciences (Wright & Mok, 2000).

Ben also collaborated extensively with students and other researchers. Many of his students have creatively extended what they learned from Ben in the areas of estimation, modeling, software, fit assessment, item banking, adaptive instrument administration, and equating to make significant contributions in their own rights. As intensely interested as Ben was in practical applications of measurement theories, methods, and software, it is not surprising that his collaborations resulted in foundational contributions in a number of fields. Similarly, his longstanding interest in the history and philosophy of science no doubt encouraged the several of his students who have taken up studies in these areas.

These contributions, along with Ben's teaching, publishing, and organizing, facilitated access to fundamental measurement across education, psychology, and the social sciences on a broad scale. Without Ben's multifaceted series of advances, it is highly unlikely that the hundredfold increase in Rasch publications over the last 40 years,[1] would have taken place. Even a cursory and sweeping glance at the scope of Ben's contributions suggests that his work is contributing to the definition of researchers' senses of their personal professional identity, as well as to the identity of the professions themselves. Though this idea plainly is generally true, Ben's 30 years of teaching a course on the psychology of becoming a teacher suggests there may be a great deal more to learn here about his motivations, ideas, and methods.

[1] As of 27 May 2017, Google Scholar shows 77 articles citing Rasch (model or analysis or scale or measurement) in 1976, and 8,380 in 2016.

14.2 Testing in the Process of Professional Identity Development

A place to start in considering how to weave the threads of Ben's story is suggested in a passage from Bettelheim (Fisher, 1991a, 2002), with whom Ben worked in the Sonia Shankman Orthogenic School at the University of Chicago in the 1950s:

> …a self, if it is not to wither away, must forever be testing itself against the nonself in a process of active assertion…. Testing implies both respect and consideration for what we test ourselves against. Otherwise it becomes not a test of self, but of something entirely different, perhaps of brute force.
>
> As a matter of fact, what a person selects as a testing ground is most indicative of the nature and quality of the self. A passive yielding to certain experiences can be a much more subtle testing of the self against the nonself than meeting it aggressively. Success is then not a question of how unchanged the self emerges from the test nor how much it has bent the nonself to its will, but how enriched it became in the process (Bettelheim, 1967, p. 81; quoted in Zaner, 1981, p. 188).

Zaner (1981, p. 188) expands upon Bettelheim's theme, saying that the

> enrichment of self must be understood not as a mere playful metaphor, but a rigorously descriptive concept. Such enrichment is a continuous, simultaneous process in which one enhances the other philosophically. We note that what is at stake is a continually ongoing, internally rhythmed and always precarious mutuality.

This sense of the way that self-development proceeds via tests of the self against others in a process of active assertion raises a question concerning Ben Wright's transitions from physics to psychology to statistics to measurement, namely: Was there something about both his psychology training and his physics training that prepared him to recognize the value of what Rasch said in 1960? Was there something that helped him not only to recognize but also to grasp and tenaciously pursue the implications of what Rasch said?

With regard to physics, the answer would seem to be clearly the full union of mathematics and measurement, that substantive integration of qualitative and quantitative data and methods in invariant relationships, characteristic of the natural sciences (Roche, 1998). Ben was intimately familiar with this capacity for the transparent and transferable identification of objects of study, and for the accumulation of knowledge. Surely Ben was intrigued by Rasch's (1960, pp. 110–115) explicit formulation of his model for reading measurement from Maxwell's presentation of Newton's second law. But Bettelheim's sense of self-other testing suggests another, psychological "something," the factor of mutuality that is as crucial to successful measurement as it is to the development and maturation of the self.

In 1955, Bettelheim and Wright co-authored a paper called "Staff Development in a Treatment Institution" and, in 1957, one titled "Professional Identity and Personal Rewards in Teaching" (Bettelheim & Wright, 1955; Wright & Bettelheim, 1957; also see Wright, 1954, 1959a, 1961b; Wright & Sherman, 1963; Wright & Tuska, 1967, 1968). These and other articles authored by Ben indicate that he and Bettelheim overtly conceived of teaching as a continuous, simultaneous process of

mutual self-enhancement through testing of the self against the other in the training of students. They support the hypotheses that Wright placed great importance on the issues raised in Zaner's quote from Bettelheim, that Wright had focused on these issues under Bettelheim's tutelage before the quote on self-other testing from "The Empty Fortress" was published, and that he continued doing so over the course of his life (Bouchard & Wright, 1997; Wright & Yonke, 1989). Closer consideration of Wright's own account of the discordances and concordances he experienced while plotting his life story shows how he integrated disparate events into a dynamic identity.

14.3 From the Individual to the Social in Professional Identity Development

Though he makes no overt reference to his work with Bettelheim, in an autobiographical account, Wright (1988b, p. 25; also see Wright, 2005, p. xi) tells us he "went in search of life" after his early "career led to an identity confusion." While pursuing a Ph.D. in physics doing almost nothing but measuring, and with his first publication, co-authored with Charles Townes (Townes, Merritt, & Wright, 1948) just out or in press, one spring day in 1948 Ben decided to seek out something livelier, more human. After exploring possibilities in English and history, he wound up in psychology, and as a consultant doing factor analyses for Chicago marketing firms. Ben's second publication, and his first as sole author (Wright, 1954), reports the results of a factor analysis, as do several other papers that emerged in the years just following (Wright, 1957; Wright & Evitts, 1963; Wright & Gardner, 1960; Wright, Loomis, & Meyer, 1963).

In Ben's (Wright, 1988b, p. 26; also see Wright, 1998, p. 20) account, the contrast between the stable, interpretable results of measurement in physics and the unstable, uninterpretable results of factor analysis put him in "considerable distress," made him feel "like a con man one step ahead of the Sheriff," and "like a crook." His self-described identity confusion resulted from the apparent incompatibility of his scientific and human values, since it seemed impossible to reconcile the physicist's demand for meaningful measurement with the psychologist's search for meaningful relationships with others.

In addition to being dissatisfied with factor analysis as a method, in the years just before he met Rasch, Ben had upset his students and faculty colleagues with his criticisms of the statistics textbook assigned for use in his initial teaching assignment in the Department of Education at the University of Chicago (Linacre, 1998, pp. 23–24). He is also in print (Wright, 1959b), before meeting Rasch, with critical comments concerning the "one-sided" conception of intelligence enacted in testing, and the mechanical sterility it imposes on children. In a paper remarkable for what it still has to teach us today concerning the integration of objective and subjective approaches to learning, Wright (1958, p. 368) asks, "What is a measurement? What

is a variable?" These are the questions, of course, that Rasch did much to help Ben answer, so their explicit articulation in print 2 years before he met Rasch is highly significant.

Wright (1961a) goes into more detail on his critical concerns with testing in a paper published in *The American Journal of Psychology*. In this brief note, he ranks items by the mean differences in two sets of ratings, foreshadowing his later interpretation of item difficulty hierarchies as evidence of construct validity (Wright, 1997, pp. 43–44; Wright & Masters, 1982, pp. 12–15, 90–94; Wright & Stone, 1979, pp. 83–93; Stone, Wright, & Stenner, 1999). Wright also emphasizes in his critique the importance of precision estimates in the interpretation of results. He imputed the variance and standard error omitted by the article's authors, en route to offering an alternative perspective on the likelihood the differences were statistically and substantively significant.

Most tellingly, in an extended and thoughtful consideration of what learning is and how the study of it might be improved, Wright (1958; see Appendix C) contrasted scientifically objective and psychoanalytically subjective approaches. To set the stage, Wright (1958, p. 366) recounts the history of humanity's progressive decenterings: being removed from the center of the solar system by Copernicus, from the crown of creation by Darwin, and from control of its own psyche by Freud (on this point, also see Ricoeur, 1970, pp. 277, 426). Wright (1958) then observes:

> Objectivity may be the royal road to reliable knowledge about the external world. But when we are trying to understand ourselves and how we learn, scientific objectivity does not seem to be enough. Perhaps we need to embark on another road, one that is more subjective. (p. 368)
>
> Occasionally the unavoidable impact of subjectivity in research is explicitly recognized. But then the influence is usually acknowledged only as a source of error. Efforts are focused on trying to rid the experiment of its subjective aspects in order to approach the hopefully scientific goal of objectivity. But these efforts at objectivity sanforise right out of the research the very data that, it seems to me, are most likely to help me out of my dilemma. Instead of trying our best to get rid of the subjective aspects of our research, we might better try our best to harness our subjective experience in a way that would allow us to sort out and make the most of its contribution. (p. 369)

Wright proceeds from here to describe an objective and two subjective approaches to the study of learning. He defines objectivity in physical terms, and while he considers animal learning investigations and classroom test scores objective and a good beginning, he says they both "barely scratch the surface of what we want to know" (p. 369). Wright points out that researchers can usually find ways to agree on the quality of the evidence produced by subjective approaches, but that agreement is not commonly deemed sufficient for the label "objective," and he wants to draw attention to their subjectivity, so that is what he calls them.

His first subjective approach concerns the emotional impact on the observer of the child engaged in learning. Attending to these feelings can add important information to an assessment. The second subjective approach "calls for a special kind of inner act," an empathic identification with another person, a student, for instance, that amounts to an application of the Golden Rule (treating others as we would like to be treated), though Wright does not refer to it as such. The goal here is to entertain

the perspective a student might exhibit in a particular behavior, and to use any insight gained "to plan a course of action that includes a feeling for what moves the child" (Wright, 1958, p. 371). Though he does not bring them up here, "a feeling for what moves the child" sounds quite like an intuition of the developmental sequences Wright later discerned in Rasch-calibrated item hierarchies. In the same way, "to plan a course of action" incorporating the scaling evidence of a child's developmental momentum and direction sets the stage for situating the learning progression defined by the items within the curriculum, as has become the focus of formative assessment, instructionally-embedded assessment, and integrated instruction and assessment over the last 20 years and more (Fisher, 2013; Wilson, 2009).

Wright (1958), not having Rasch's models, measurement theory, or experimental results at hand to work from at the time he was conceiving his personal approach to learning, acknowledges that these subjective approaches involve difficulties, of course. The most troublesome issue that emerges is the previously mentioned problem "that we are all as much subject to, as we are masters of, our own state of mind." That is, one might well project unwanted features of her or his own makeup on others, or deny or repress those features, with negative consequences for research and practice. These difficulties can be overcome in his personal approach to learning, Wright (1958, p. 372) notes, by sharing observations with others, obtaining their feedback, and by complementing the subjective approaches with objective information.

Wright's work in educational measurement effectively blends one's subjective feeling for what moves the child with objective evidence of the direction and pace of that movement. Subconscious projections, repressions, and denials must confront not only the facts of the observed assessment results, but must also contend with results produced from multiple assessments and explained from other perspectives. Lasting value can be expected to result from the convergence and divergence of these multiple sets of results, as in fact has been increasingly recognized in the tangible learning gains produced by formative assessment feedback (Black, Wilson, & Yao, 2011; Hattie, 2008).

But what Wright accomplished in his formulation of a personal approach to learning amounts to nothing less than an independent beginning at what has been called a simultaneously objective and subjective "joint epistemic project addressing the historically changing and mutually conditioning relation of 'inside' and 'outside' knowledge" (Galison, 2008, p. 293). Cycling between subjective reflections and objective observations in a relational ontology is a form of the dialectic of belonging and distanciation described by Ricoeur (1976, p. 79) as a circularly related process of guess and validation. Understanding begins as a guess, but is transformed when it encounters the objective text of, for instance, a student's response to a question. Wright's personal approach to this dialectic stands paradigmatically apart from the modern, Cartesian worldview, just as his measurement philosophy and methods stand apart from the usual juxtapositions of statistical and qualitative data in mixed methods research (Fisher & Stenner, 2011a).

The crux of the matter is that Descartes failed to account for "the circle in which he was involved when he presupposed … the possibility of inferences transcending

the ego, when this possibility, after all, was supposed to be established only through this proof" (Husserl, 1970, p. 90). The simultaneous projection and taking up of the possibility that inferences could transcend the ego was Descartes' brilliant, but "ontologically defective" (Heidegger, 1962, p. 128), metaphysical expression of Galileo's similarly "ambiguous genius [that], in uncovering the world as applied mathematics, covers it over again as a work of consciousness" (Ricoeur, 1967, p. 163; Husserl, 1970, pp. 23–59; Burtt, 1954, p. 204).

How do we then include the possibility that inferences will transcend the ego in the proof? How do we uncover the world of human cognition and behavior as applied mathematics without covering it over again as a work of consciousness? Gadamer (1989, p. 104) provides a hint when he points to the mode of being of play as an important methodological clue. The fact that light, waves, animals, etc. all play, really play, and that humans too play, is the route toward the conceptualization of the subject that will engender a conception of learning processes that applies as much to the play of natural forces as it does to the play of psychosocial forces (Fisher, 2017).

That is, circling the presupposition of inferences transcending the ego back on an instance of such an inference, we necessarily arrive at the "I am," a moment in identity development. But instead of building out from here in a linear logic a la Descartes, we must instead allow this internal dialectic of belonging and distanciation to resonate and vibrate with the beating heart of rhythmically emerging patterns of interactions with the things and others around us. In so doing, "thinking thinks itself" and we take "cognizance of that which we already have," as Heidegger (1967, p. 104) put it. In Wright's personal approach to learning, what teachers already have is a conceptual model of the kind of information needed to develop a feeling for what moves the child. As argued by Fisher (2003a, 2003b, 2004, 2010a; Fisher & Stenner, 2011a), given objective information in the form of answers to questions and feedback from the student and others in the environment, teachers enact what Heidegger's student, Gadamer (1989, p. 367), calls the "art of testing," which is an "art of questioning." Questioning of this kind focuses "not on trying to discover the weakness of what is said, but in bringing out its real strength" (Gadamer, 1989, p. 367). Testing in the context of formative assessment (Black, Wilson, & Yao, 2011; Fisher, 2013; Wilson, 2009) is increasingly recognized as a tool for realizing exactly this purpose: helping the teacher locate the student relative to the already known curriculum and learning progression, and helping the student connect with the positive value of what is already known so as to employ it in gaining new knowledge.

Teaching thus is fundamentally a transactional process of guiding students to the discovery of what they already have (Dewey & Bentley, 1949; Romer, 2013). No learning occurs if the student merely takes what is given in instruction in a way that does not involve experiencing what is taken in terms of what is already known (Heidegger, 1967, p. 75). Even in any mundane everyday task, we organize, form, and live out our self-identities relationally, through processes of dialogical interactions with the world that teach us what is going on right now in terms that must connect with what we already know (Overton, 2015).

Wright (1958, p. 369) effectively formulates his own relational ontology when he expresses the concern that even when the unavoidable impact of subjectivity is recognized in research, that usually happens only to acknowledge it as a source of error to be removed. Wright fears that this eliminates precisely the resources needed to counter what he feels are the mechanistic and sterile consequences of purely objective methods. Latour (2004, p. 219), echoing Wright, offers some observations that become salient here:

> …neither distance nor empathy defines well-articulated science. You may fail to register the counter-questioning of those you interrogate, either because you are too distanced or because you are drowning them in your own empathy. Distance and empathy, to be useful, have to be subservient to this other touchstone: do they help maximize the occasion for the phenomenon at hand to raise its own questions against the original intentions of the investigator—including of course the generous 'empathic' intentions? It must be clear, according to this formulation, that abstaining from biases and prejudices is a very poor way of handling a protocol. To the contrary, one must have as many prejudices, biases as possible, to put them at risk in the setting and provide occasions of manipulation for the entities to show their mettle. It is not passion, nor theories, nor preconceptions that are in themselves bad, they only become so when they do not provide occasions for the phenomena to differ.

In Gadamer's (1981) words, "the fruitfulness of scientific questioning is defined in an adequate manner if it is really open to answers in the sense that experience can refuse the anticipated confirmation" (p. 164). A primary goal for educational research and practice in this paradigm becomes conceiving, gestating, midwifing, and embodying subjective understanding and objective explanations together in the material form of instruments that are calibrated and applied via processes that put prejudices and biases at risk of being refuted. The anticipated confirmation may be refused by data examined in relation to hypotheses, or by an inability to devise a theory capable of explaining the observations well enough to predict new ones.

Following in the path of Latour's (1987, pp. 247–257) focus on metrologies, instruments in this paradigm also (a) are written and read in the shared languages of common metrics, (b) are interpreted primarily in qualitative, not quantitative, terms, (c) always include indications of uncertainty, and (d) are distributed throughout cognitive ecologies for application and interpretation at the point of use (Fisher & Stenner, 2017). Scientific learning, like classroom learning, takes place in terms of what is already known. When we lack instruments embodying what is already known about a construct's invariant structure, we lack a common language deploying shared concepts and terms, and we have no medium through which we can expand our networks of interrelated experiences. The failure to create these media in psychology and the social sciences is all the more tragic given that scientific innovation depends on unobstructed flows of information. Shared standards are widely recognized for their value in creating efficient economic, commercial, and scientific markets (Miller & O'Leary, 2007). Situating Wright's personal approach to learning and contributions to measurement in this context shows that his work is likely to have far reaching philosophical, methodological, and practical consequences.

On a more immediate level, Wright's personal approach to learning leads to a new appreciation for the interplay between participant and vicarious learning. In this context, Wright also recognizes the role of mistakes in learning: "While we emphasize getting the right answer in our schools and on our tests, for example, we have all had the experience of learning more from our mistakes than from our right answers." This concern for the value of mistakes in learning is in tune with Bettelheim's sense of self-other testing, and also comes up in Wright's introduction to Nielsen (1968), where Wright (1968a, pp. 13–14) advocates "the recognition and acceptance of failure as a real curriculum."

In this, he presages Carol Dweck's (2006) research into fixed and growth mindsets, saying "Nothing like success so bars the way to coming to grips with failure, to living with it, to digesting it, to making the most of it" (Wright, 1968a, p. 14). Ben's later work on partial credit Rasch models (Wright & Masters, 1982; Masters, 1982, 1984) and on kidmaps (Wright, Mead, & Ludlow, 1980; Masters, 1994; Mead, 2009; Chien, Linacre, & Wang, 2011) focuses on the quantitative and qualitative display of both expected and unexpected failures for individual students. Building these tools for learning into accountability methods, Wright (1977, p. 108) also later connected local classroom and national assessment needs, foreshadowing more recent attention to situating unique individual item response patterns within formative classroom assessments that are coherent relative to high-stakes tests (Wilson, 2004; National Research Council, 2006; Gorin & Mislevy, 2013; Fisher, Oon, & Benson, 2017).

But, Wright (1958, p. 375) continues, "the main topic to which a personal approach to learning leads is interpersonal relations and their central role in learning, including the vicissitudes of that profound and mystifying phenomenon called 'identification.'" Ben is not cited for his contributions in this regard in a recent publication (Olitsky, Flohr, Gardner, & Billups, 2010) on student identity formation, but this article echoes several of his early themes. These include criticisms of standardized testing (Wright, 1959b); his focus on unique, local, qualitative aspects of personalized assessment and instruction (Wright, 1958); his inclusion of students as participants in multifaceted dialogues and not merely as recipients of transferred knowledge (Wright, 1954, 1958; Wright & Bettelheim, 1957; Wright & Tuska, 1965b); and his revival of proposals for students' mutual instruction (Wright, 1960; see Appendix B). Sisson's (2016) recent citations of Ben's early work on teacher identity focus on its emphasis on teachers' personal histories with their own teachers as models for professional behavior, remarking on the need for more attention to contextual issues. Contextual issues were, in fact, taken up in an American Psychological Association conference presentation by Wright and Tuska (1965a).

This array of concerns foreshadows Ben's later work in measurement, where he found ways to integrate subjective experience and objective criteria for qualitatively meaningful and reproducible quantification. Even before his ideas developed very far in the direction indicated by Rasch's models, and before he adopted those models in his own work, Wright and Evitts (1963) used principal components factor analysis to investigate the "unidimensional structural properties of objective attributes," citing Lumsden's (1961) paper on this topic.

But in Rasch's ideas on measurement, Wright saw how to retain what is most valuable about subjective identification and empathy by giving one's consciousness over to the play of the object of the question and answer dialogue. A desire to understand posits best guesses from past experience as hypotheses to be tested against observations, instantiating the dialectic of belonging and distanciation. As he (Wright, 1977, p. 97) later put it,

> When a person tries to answer a test item the situation is potentially complicated. Many forces influence the outcome—too many to be named in a workable theory of the person's response. To arrive at a workable position, we must invent a simple conception of what we are willing to suppose happens, do our best to write items and test persons so that their interaction is governed by this conception and then impose its statistical consequences upon the data to see if the invention can be made useful.

Ben saw that subjectively submitting to the repeatable and reproducible production of invariant profiles across tests and samples of students led to objective results not predetermined in a one-sided way, but which emerged from sympathetically caring for the unity and sameness of the discourse, in accord with Gadamer's (1991, p. 61) sense of "the first concern of all dialogical and dialectical inquiry." Wright's goal was to write items and administer them in ways that allow, so far as possible, the question and answer interaction to be governed by a simple conception, an hypothesis, of what our informed opinion leads us to think might happen.

Wright saw that a student's measure estimated in the context of a Rasch model offers an opportunity for empathically identifying with that student, and for planning a course of action based on a feeling for what moves the child. Student measures are positioned along a learning progression or developmental sequence defined by those students' experiences of how difficult it is to respond correctly to the questions asked. The invariance of the item hierarchy, within the range of uncertainty and taking unexpected inconsistencies into account, tells a story about how learning progresses from what is already known to what is not yet known. Easier items establish areas of content that are typically understood before harder items can be answered correctly.

This story is not a mere enumeration, a simple count of correct answers used to establish the order of things. It entails instead a complex combination of theory and evidence applied to the validation of the construct, and the inferences and consequences drawn from it (among many others, see Bond & Fox, 2015; Dawson, Fischer, & Stein, 2006; Engelhard, 2012; Fisher & Stenner, 2011a, 2016; Stenner, Fisher, Stone, & Burdick, 2013; Stone, Wright, & Stenner, 1999; Wilson, 2004, 2005). Furthermore, vertical scales measuring development over the course of several years are not, of course, calibrated on complete data (responses from every student on every item). Following on the early work of Wright (Wright & Douglas, 1975; Wright & Bell, 1984) and his student, Choppin (1968, 1976), it is increasingly common for individualized student assessments to be constructed from precalibrated item banks. In these applications, many students will obtain the same count correct score by answering items of different difficulties, and they will, then, also have measures indicating their different abilities.

It may seem that this kind of evidence-based approach to seeing what to teach next, how to leverage what the student already knows to move her or him via an instructional plan, is more logical than empathic. But many teachers react negatively to test questions that turn out to be far more difficult than they expect them to be. The common inclination is to blame students for not mastering the material as it was presented, or to blame the quality of the teaching or the textbook. Demonstrations of the generality of the item hierarchy, explanatory theories illustrating the structure of the learning progression scaffolding, and broader dissemination of integrated assessment and instruction curricula, will perhaps be needed to overcome some teachers' lack of empathic identification with students' learning experiences.

Though more attention must be paid to the inferential differences between cross-sectional and longitudinal scaling (Williamson, Fitzgerald, & Stenner, 2014), the story told by the item hierarchy is true in general even though it is not specifically true of anyone student in particular. As Rasch (1960, pp. 37–38; 1973/2011; also see Box, 1979, p. 202) vigorously asserted, the point is not the truth of the story told, but its pragmatic utility. Neither Newton's laws nor the Pythagorean theorem are strictly true (Cartwright, 1983), but their value in organizing experience and managing life are well established. Increasing attention to issues of information coherence in developmentally, horizontally, and vertically aligned assessments (Wilson, 2004; National Research Council, 2006, p. 26; Gorin & Mislevy, 2013) complements other recent work on the potential for forms of metrological traceability (Fisher, 2009, 2012; Fisher & Stenner, 2011b, 2016; Mari & Wilson, 2014; Pendrill & Fisher, 2015; Wilson, 2013b) based on Rasch's models for measurement. In both cases, researchers are exploring opportunities for creating new media for writing identity narratives expressing the fulfilment of broader, deeper, and more fulfilling life choices.

Wright's (1958) personal approach to overcoming the limitations of subjectivity by sharing observations with others and complementing subjective approaches with objective information was effectively instantiated in his approach to measurement. The potential weaknesses of subjectivity were mitigated by the process of embodying hypothesized suppositions and empathic identifications in the questions asked, and then carefully studying the responses in a way capable of revealing intrusive departures from the overarching pattern of invariance. Subjective projections, denials, or repressions imposed by the researcher in the effort to identify and empathize with students in their learning experience would potentially be made visible in the statistical and graphical evaluation of the response consistencies relative to the model, and would so be made actionable by means of this process.

Ben's nondualistic integrated subject-object approach explicitly includes the researcher's subjectivity on par with the research subjects' experiences. This shared subjectivity is drawn into a dialogical recounting of a narrative as invariant, objective, and reliably reproducible as the uncertainty estimates, consistency evaluations, validity evidence, and explanatory theory allow. The autonomy of the data text stands relative to the meaning of the measured construct. The object of the dialogue unfolds by means of the playful interaction through which it reveals itself. This is quite in tune with Gadamer's (1989, pp. 101–134) sense of play as the basic clue to

a method of fused subject-object horizons, and his (1989, p. 367) sense of conversation as an art of testing that involves a Socratic art of questioning (Fisher, 2003a, 2003b, 2004).

Thus, after some time in the awkward position of feeling stuck with factor analytic and testing methods he did not respect, and having articulated his own quite sophisticated personal approach to learning, in 1960 Wright met Georg Rasch (see Wright, 1998, for his biography of Rasch, and an account of their personal relationship). He joked (Wright, 1988b, p. 27) that he could then "stop going to the psychoanalyst to have my schizophrenia mended week by week." What Rasch (1960, 1961, 1977) offered was an experimental approach to evaluating the possibility of stable results that would remain invariant over samples of examinees and test items.

Rather than factor loadings that would change across samples from week to week, Rasch's models made it possible to think in terms of scale metrics that would stay the same to the extent that the experimental results and explanatory theory supported them. The end result is a capacity to integrate discordances and concordances into variations on an invariant theme. Individual students formatively assessed in classrooms could have tests adapted to their abilities (Wright & Douglas, 1975), and instruction custom tailored via kidmaps (Wright, Mead, & Ludlow, 1980). Teachers, researchers and practitioners making use of the models to guide the construction, implementation, and interpretation of the student measures would be better able to create coherent narratives of both their students' and their own performances and identities as learners.

Ben does not mention them in his account of how he resolved his "identity confusion," but he was well aware of the ways in which the developmental process of individuation requires a decisive break with a key mentor (Wright, 1959a, pp. 365–366; Wright & Yonke, 1989). This break is a separation of the student's identity from the teacher's, and emerges as a result of the process of actively asserting, and so testing, the self against the other.

In Ben's (Wright, 1988b, p. 27) own account he made an ineffectual step in the direction of establishing his own professional identity in 1964 when he contradicted Rasch by incorporating an item discrimination parameter into software he was writing with Bruce Choppin. They had written logarithmic, pairwise, and recursive symmetric functions (conditional) programs that all gave the same results. Then, observing that the fit lines varied in slope, Ben thought, "Let's estimate the slopes too." Rasch "was very much against this bright idea." This step was ineffectual as a path toward creating an independent professional identity for Ben most obviously simply because he "couldn't get it to converge." That is, this approach negated the value he had found in Rasch's models relevant to the invariant stability of the results across samples.

As Ben's consultation with a mathematician (Adrian Albert) showed, the extra item parameter made it impossible to maintain a single item order (within the limits of uncertainty and precision) over examinees, and vice versa. Arbitrary starting values were often needed to achieve convergence, and results would vary across the different starting values chosen (Stocking, 1989). Wright (1977, p. 103; 1984, 1997, pp. 40–44) came to see that "additional parameters like these…wreak havoc with

the logic and practice of measurement." This inability to arrive at stable results is a function of over-parameterization, referred to in recent years as a matter of model identifiability (Verhelst & Glas, 1995, pp. 235–236; San Martin, Gonzalez, & Tuerlinckx, 2015).

Had Ben remained invested in this particular way of distinguishing his professional identity from Rasch's, he would have been no better off than he had been before he met Rasch, when he had felt like a crook or a schizophrenic. He would not have been any better off because the extra item parameter introduced the same kind of instability that his factor analysis results had. Both the factor loadings and the model with the extra item parameter similarly lacked a clear approach to explaining or correcting the instabilities they revealed, which is one of the strengths of Rasch's separability theorem and concept of specific objectivity.

To have written the factor analysis and extra item parameter models into his life story, Ben would have had to be satisfied with a narrative of multiple separate variations, instead of imaginative variations on an invariant, in Ricoeur's (1991b, p. 196) terms. This would have put Ben in the position of compromising the values he had worked into his personal approach to learning. He would have had to relinquish having a clear criterion for knowing if and when his own projections, denials, and repressions were coloring his interpretation of himself and others. (We will save for another time consideration of any possible relevance here of Ben's multiple published studies (Wright & Gardner, 1960, for instance) in the period of 1960–1962 on the meaning and psychological effects of color.)

Wright let go of the second item parameter, which became a primary feature of Item Response Theory models (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Van der Linden & Hambleton, 1997). Ben made a cleaner break a few years later with the development of the unconditional maximum likelihood (UCON, now referred to as Joint Maximum Likelihood Estimation, or JMLE) algorithm (Wright & Panchapakesan, 1969; Wright & Douglas, 1977; Wright, 1988a) for estimating model parameters. Rasch opposed this move as well but it retained the connection with invariance, parameter separation, and sufficient statistics in a way that the earlier two parameter program did not. In Wright's (Andrich, 1995, p. 4) own words, his UCON work

> was an important point in our [Ben and Rasch's] relationship because at that moment he and I separated a little bit. Up until then, as far as he was concerned, I was doing everything exactly the way he told me. But UCON was a new something that I did on my own, not to his liking, which seemed to me plainly convenient, practical and useful. So it was a point in our work where I was becoming myself, in spite of, indeed, against his wishes. We continued to be good friends. But from that summer of 1967, there was that bit of difference between us.

By taking this step, Ben took responsibility for advancing his own ideas and innovations in a direction not specifically foreseen or supported by his teacher. As Ben already well knew from his work with Bettelheim, this meant he had completed a significant stage in the development of his own identity as a professional. His explicit awareness of the importance of this step raises the question as to whether he might have tried deliberately to provoke others into taking it.

Ben alternated contributions concerning improved access to measurement with provocations to measure better, and to think more clearly, as an intrinsic part of professional life. Rasch's models integrate information about individuals with that of the populations to which they belong at the same time that they abstract that information from them. Similarly, Wright simultaneously supported the professional development of both individuals and populations by making measurement more accessible, and by provoking others into overtly testing and asserting the validity of their own measurement innovations and contributions.

In my own case, for instance, I had the great fortune of discovering on my first day in Ben's classroom concepts and tools that I had previously thought I was going to have to invent. But this revelation of open access to what I recognized to be of great value was soon (within 2 or 3 weeks) countered by Ben's flat dismissal of my approach to the language of measurement theory. I wrote an impassioned paper explaining my position, was pleasantly surprised that Ben warmly embraced my point of view, and that he additionally showed a plain respect for my having pushed back at him in an assertion of my independent identity.

In other cases, we see the extent to which Ben's students, and the colleagues he influenced, have developed strong professional identities and so have taken up positions of leadership:

- in further model developments (e.g., Adams, Andrich, Karabatsos, Linacre, Masters, Wilson),
- in the exploration of new estimation methods (e.g., Adams, Karabatsos, Wilson),
- in new ways of assessing model fit (e.g., Douglas, Mead, Ludlow, R. Smith,),
- in making measures richer in meaning and interpretive applicability (e.g., Linacre, Masters, Stenner, M. Stone, Woodcock, Wilson),
- in innovations in instrument administration (e.g., Bergstrom, Douglas, Gershon, Lunz),
- in schools of education (Andrich, Bond, Engelhard, Green, Karabatsos, Ludlow, Mok, Myford, E. Smith, Wilson),
- in educational research organizations (e.g., Adams, Bergstrom, Bontempo, Masters, Mislevy, Myford, Schulz, R. Smith, Woodcock),
- in professional certification standard setting (e.g., Bergstrom, Lunz, Shen, R. Smith, G. Stone, Surges Tatum, Wendt),
- in their own commercial enterprises (e.g., Bezruzcko, Bontempo, Gershon, Lunz, Stenner, Surges Tatum),
- in software (e.g., Adams, Andrich, Douglas, Linacre, Ludlow, Moulton, Schulz, R. Smith, Wilson), and
- as early adopters of Ben's innovations in a wide variety of fields, including
  - professional competence evaluation and certification,
  - coatings technologies,
  - developmental psychology,
  - physical therapy,
  - physical medicine and rehabilitation,
  - internal medicine,

  - sport psychology,
  - nursing,
  - occupational therapy,
  - writing assessments,
  - judged performances,
  - psychiatry, and many others;
- and in numerous countries around the world, from Kuwait to Korea, and from Malaysia to South Africa.

A pressing question is how we as individuals and as a field are now to respond to the access and provocations of Ben's work. Information relevant to formulating this response emerges from historians' and social scientists' examinations of the vital role played by interactive system effects in the history of science (Lenoir, 1997; Galison, 1997; Golinksi, 2012; Latour, 1987, 2005; Nersessian, 2006; Fisher & Wilson, 2015; Woolley & Fuchs, 2011). It appears that the success of science follows less from a compelling consensus on data, theory, and instruments than it does on a balance between dissonance and harmony across its various communities. Could Ben's "imaginative variations on an invariant" and use of stochastic measurement as a medium of identity formation reach beyond the development of individual professional identities to the development of the identities of professions?

The most obvious place to look begins from the rampant imbalances between divergent and convergent perspectives in psychology and the social sciences. Divergence and a lack of care for invariant constructs and measures is the order of the day, leading some to speak of methodological pathologies in psychology (Michell, 2000). This issue is distinct from the productive disunity observed across communities of experimentalists, theoreticians, and instrument makers by Galison (1997), as it also is distinct from the divergent and convergent efforts and discourses observed by Woolley and Fuchs (2011) in their examination of collective intelligence in science. Indeed, in both of these studies, metrology and common languages play key roles in making divergence and disunity salient.

It is then important to recognize that, far from presenting Wright as singlehandedly founding a new measurement discipline, or even as cultivating a community of researchers sharing a common paradigm, "the multidimensional linkages and exclusions of and between different discursive practices required for the creation of a discipline exceed the power of individuals to engineer and orchestrate" (Lenoir, 1997, p. 52). Insofar as Wright's wide-ranging contributions to measurement theory and practice (for a representative sample, see Fisher & Wright, 1994) contribute to new directions in science, it will be because system effects emerged from the interactions of dispersed individuals with different and only loosely related agendas (Fisher & Cavanagh, 2016; Fisher & Wilson, 2015).

So, without denying the need for healthy disagreement and conflicting perspectives, in light of Wright's legacy of longstanding methods and results demonstrating how to calibrate and maintain invariant metrics across samples and instruments, it is way past time to recognize that the persistent practice of measuring the same construct in different units is perversely self-defeating and counterproductive. After

all, to what extent is psychology, sociology, or any other *ology* actually fulfilling its mission as an effective manner of expressing a particular field of meanings if its logos (putative proportionate rationality) remains blatantly dependent on the particular persons and phrasings of the questions and answers embodying the conversation?

In other words, to what extent does a field of study actually have a professional identity if its objects and subjects are not clearly expressed and distinct from those of other fields? Is a field's identity coherent if its variations on an invariant do not plot a meaningful story? In Latour's (2004, p. 218) terms,

> If there is a physio-logy, a psycho-logy, a socio-logy, a glacio-logy, an ethno-graphy, a geography, etc., it is because there exist laboratory settings where propositions can be articulated in a non-redundant fashion. As the etymology of those disciplines nicely indicates, talking and writing is not a property of scientists uttering statements about mute entities of the world, but a property of the well-articulated propositions themselves, of whole disciplines.

There are many expressions of the opinion that fields of study are only as scientific as they are mathematical (for instance, Kant, 1970, p. 7), not all of which automatically assume the mathematical to be numerical and quantitative (Heidegger, 1967, p. 68; Kisiel, 1973; Fisher, 2003a, 2010a, pp. 12–14). Ben (Wright, 1968c, 1977, 1984, 1985) always strove to make clear that the mathematical means quantitative far less than it implies a rigorous qualitative independence of figure from meaning. Indeed, quantification is neither necessary nor sufficient for measurement (Mari, Maul, Torres Irribarra, & Wilson, 2016), and the meaningfulness of well-articulated propositions is a function of the same invariance as found in measurement (Mundy, 1986; Fisher, 2003a, 2003b, 2004, 2010a).

This point concerning the qualitative aspects of quantification was already being made in psychometrics before Ben came on the scene (Loevinger, 1947; Guttman, 1994, p. 82; Thurstone, 1959, pp. 9–10). It has become of increasing interest as construct validity and meaningful inference have become matters of central concern (Dawson, et al., 2006; Mundy, 1986; Fisher & Cavanagh, 2016; Fisher & Stenner, 2011a; Wilson, 2005). Rasch's (1960, 1961, 1977) separability theorem provides the basis for hypothesizing and experimentally testing that independence. That theorem then stands to play a key role in the development of professions' identities, since these depend no less than individual identities on the way they test themselves against others. Science and scholarly learning aspire to ideals of cumulative results and universal access to those results, which requires negotiating disjunctive discontinuities. Though the stories we tell may well abruptly change direction or exhibit frustration with the way every effort at wringing meaning from events seems to fail, the medium is the message. In trying to tell a story at all, there is hope for the possibility of making sense of, and sharing, life, as Ricoeur (1974c) contends in his contrast of violence and language.

14.4 Towards a New Art and Science of Living Meaning

Far from recommending or attempting a shallow imitation in psychology of physics, Wright rethinks method in an original way, integrating objectivity and subjectivity in a compelling personal approach to learning. In so doing, his work anticipates and parallels more recent investigations showing how everyday model-based reasoning serves as a basis for the formation of new concepts in scientific research (Nersessian, 2006). It is telling that Maxwell's method of analogy plays a key role in both the historical accounts provided by Nersessian (2002, 2006) and in Rasch's (1960, pp. 110–115) formulation of his models (Fisher, 2010b). Though modeling in psychology and the social sciences will have to become much more than mere data analysis to fulfill its potential (Stenner, et al., 2013; Wilson, 2013a), perhaps richer integrations of data with theory and instrumentation will be motivated by more widespread appreciation of the importance of the fact that "a significant segment of history and philosophy of science now gives models and modeling pride of place among scientific tools and practices" (Nersessian, 2008, p. 204). Wright's role as a leader in model development and application suggests his work will play key roles in new discoveries for some time.

And so it may be that, in the same way that, first, harmonic and geometric studies, and later, the modern sciences, emerged from Socrates'maieutic (midwifing) tests of ideas as hypotheses, today we are witnessing the conception and birth of new forms of understanding relevant to broadly conceived, simultaneously qualitative and quantitative mathematical structures in the human sciences (Dawson, et al., 2006; Fisher, 2003a, 2003b, 2004, 2010a; Fisher & Stenner, 2011a). Ben Wright developed and established his professional identity relative to Rasch's and that of his students and colleagues. He provoked, both constructively and destructively, others to distinguish themselves from him, in turn. And he implicitly enhanced the identities of a wide range of professions by making it possible for them to better clarify their objects of inquiry, and their own status as communities of inquirers.

Ben integrated his combination of objective and subjective approaches to learning in practical measurement models, estimation methods, fit and bias assessments, software, statistics and graphical displays, and instrument and report designs. He deployed these media in the energetic mentoring of students and colleagues, the founding of professional societies and publications, and in applications across dozens of fields.

In so doing, Ben contributed significantly to the initiation of a new nondualistic, noncartesian, unmodern (Dewey, 2012) or amodern (Latour, 1993) paradigm in the history of science. His tacit grasp of the full meaning of McLuhan's phrase, "the medium is the message," led him to see the value of providing wide access to tools capable of embodying probabilistically consistent, meaningful, and interpretable accounts of who we are as individuals and communities. In so doing, Wright points toward new ways of addressing what Dewey (1954, p. 216) considered the most urgent problem of contemporary life: that the public find and identify itself. These stories are never perfectly transparent and reliable, but always entail uncertainty and

unexpected discordances. As Ben well understood, bringing these into view is often as—or even more—important and valuable as harmonious confirmations of expectations.

Ben wrote the (1968a) Introduction in Nielsen's (1968) *Lust for Learning*, a book on the philosophy of education adopted at Nielsen's New Experimental College, in Thy, Denmark. In this introduction, Ben recognizes and accepts not only the previously mentioned curriculum of failure, but also celebrates uncertainty as "the garden of creation" that "can never settle on fixed courses, never finally define the agenda for good enough instruction, [and so] is a curriculum that never stops growing" (p. 14). Even more importantly, Ben (1968a, p. 11) speaks of lust as "THE neglected human motive. It is the missing ingredient in the philosophy of education and the psychology of learning."

In saying this, Ben echoes Socrates, who said that we are enthralled with meaning in the same way a lover is captivated by the beloved. Following Plato's recounting in the *Symposium* of Socrates' story of his meeting with Diotima, we see that it "is through this extraordinary phenomenon of love that we thereby come to understand how meaning can be thought about. For in thinking about meaning, we neither fully possess the perfect form of meaning (e.g., the ideal state), nor are we totally unaware of it" (Gelven, 1984, p. 132). Human being is fundamentally and irrevocably caught up in meaning. One of the primary philosophical consequences of the quantum physics Ben studied is the realization that "we are suspended in language in such a way that we cannot say what is up and what is down" (Neils Bohr, in Petersen, 1968, p. 188). Desire for meaning and captivation with beauty together make lust for learning what it is.

Ben was invited to write the introduction to Nielsen's book because, in 1967, as part of his travels in Denmark to work and study with Georg Rasch, Wright enrolled in the New Experimental College (NEC), and participated in its courses and organization. Ben had previously visited Rasch in Denmark in 1964 and 1965. When he returned in 1967, he had expanded his scope of activities to include joining the NEC faculty. Ben continued this association for some time, as the Back Matter pages of the Spring, 1970, issues of *The School Review* and *The Elementary School Journal* both include announcements of a workshop to be given by Ben at NEC that August on the psychology of becoming a teacher.

During his visit in 1967, Ben gave a Sabbath Lecture (Wright, 1968b; see Appendix A) at NEC on Saturday, September 16, speaking on the conflict and communion of love and order. At least in part because of a study of the metaphor "love is a rose" in my dissertation, Ben gave me a copy of Nielsen's book on the NEC the week I received my PhD. On the title page he wrote: "To Bill: See page 67. from Ben, 6/10/88." This page in the lecture, as we will shortly see, indirectly expands on Ben's point in the introduction to the book that lust for learning is the missing ingredient in the philosophy of education.

Ben's Sabbath Lecture topic is a variation on the ancient theme of *ordo amoris* taken up by St. Augustine, Pascal, and philosophers like Max Scheler (1973, pp. 98–134). Pascal's words are perhaps the most well known in this genre: "The heart has its reasons, of which reason knows nothing." Essential to the reasons of the heart is that they are both individual and collective (Welten, 2016, p. 140). Ricoeur (1974a, p. 245) remarks that "…the only true creators, it seems, are those who are capable at the same time of reactivating the meaning of or the feeling for an *ordo amoris* which it is not ours to create." The meaning of or feeling for love's reasons are not ours as individuals to create because they, as Plato put it, draw us toward them with a will of their own that can often be quite at odds with our own will.

To write coherent narratives unifying past, present, and future, the human sciences need to effect a transition from description to prescription, from judgments of dead facts to judgments of living value (Ricoeur, 1992, pp. 169–171). And more than that, "a social ethic cannot spring from a system but from a paradox. It aims at two opposed things: human totality and human singularity" (Ricoeur, 1974b, p. 166). Ricoeur continues, saying, "I want both." Wright, too, wants both, as can be seen in his Sabbath Lecture. The paradox of wanting both human totality and human singularity becomes much less an apparently unresolvable contradiction when approached in terms of the hierarchical complexity of progressive developmental integrations (Dawson, 2002). Wright's work played an important role in scaling and reconciling different theoretical perspectives in this area. Much more could, and likely will, be said on Wright's lesser known work, from the pieces foreshadowing his seminal contributions to measurement introduced here, to others not yet explored (for instance, David & Wright, 1974; Levinsohn & Wright, 1976). For now I will take my leave, and allow Ben (Wright, 1968b) the last word:

> My text is the conflict of *love* and *order*. What are they? To bring this out I will oppose them and through their contrast try to clarify their distinctive characters.
>
> If love is feeling, then thought is order. If order is action and structure, sensation and flow belong to love. In human relating we think of responsiveness and reunion as loving. Then self-assertion and individuation are orderly. Communion is an expression of love. Identity is an expression of order. Order is the foundation of clarity. But love can lead to confusion. Love nourishes hope and creation, and promotes the rich experience of life. But order aims at discipline and conclusion and requires in the end the fixed settlement of death.
>
> Love and order also play their part in the practice of science. Full experience of reality is an expression of love. But that narrowing and selection of experience which becomes scientific observation is an act of order. The control and organization of observations which become theory are triumphs of order. But the response to theory which becomes understanding and insight are celebrations of love. (pp. 66–67)
>
> We are by order arranged but by love possessed. (p. 67)
>
> …the communion of love and order…is order reworked for the sake of love, and love harvested for the nourishment of order—order for love *and* love for order. If this communion has a sacrament, then it will be found in the fruitful human relation. (p. 68)

References

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

Andrich, D. (1982). An index of person separation in Latent Trait Theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives, 9*(1), 95–104.

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

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*, 449–460.

Andrich, D. (1995). Rasch and Wright: The early years (transcript of a 1981 interview with Ben Wright). In J. M. Linacre (Ed.), *Rasch measurement transactions, part 1* (pp. 1–4). Chicago: MESA Press.

Arnold, S. F. (1985). Sufficiency and invariance. *Statistics and Probability Letters, 3*, 275–279.

Bettelheim, B. (1967). *The empty fortress: Infantile autism and the birth of the self*. New York: The Free Press.

Bettelheim, B., & Wright, B. D. (1955). Staff development in a treatment institution. *The American Journal of Orthopsychiatry, 25*(4), 705–719.

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.

Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Bouchard, E., & Wright, B. D. (1997). In M. Protzel (Ed.), *Kinesthetic ventures: Informed by the work of F. M. Alexander, Stanislavski, Peirce, and Freud*. Chicago: MESA Press.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–235). New York: Academic Press.

Burtt, E. A. (1954). *The metaphysical foundations of modern physical science* . Garden City: Doubleday Anchor.

Cartwright, N. (1983). *How the laws of physics lie*. New York: Oxford University Press.

Chien, T.-W., Linacre, J. M., & Wang, W.-C. (2011). Examining student ability using KIDMAP fit statistics of Rasch analysis in Excel. In *Communications in Computer and Information Science: Vol. 201. Advances in Information Technology and Education* (pp. 578–585). Berlin: Springer.

Choppin, B. (1968). An item bank using sample-free calibration. *Nature, 219*, 870–872.

Choppin, B. (1976). Recent developments in item banking. In D. N. M. DeGruitjer & L. J. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 233–245). New York: Wiley.

David, T. G., & Wright, B. D. (1974). *Learning environments*. Chicago: University of Chicago 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., Fischer, K. W., & Stein, Z. (2006). Reconsidering qualitative and quantitative research approaches: A cognitive developmental perspective. *New Ideas in Psychology, 24*, 229–239.

Dewey, J., & Bentley, A. F. (1949). Knowing and the known. Boston: Beacon Press.

Dewey, J. (1954). *The public and its problems*. Athens: Swallow Press, Ohio University Press.

Dewey, J. (2012). Unmodern philosophy and modern philosophy. In P. Deen (Ed.). Carbondale: Southern Illinois University Press.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York: Random House.

Engelhard, G., Jr. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement, 8*(1), 21–38.

Engelhard, G., Jr. (1994). Historical views of the concept of invariance in measurement theory. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 73–99). Norwood: Ablex.

Engelhard, G., Jr. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge Academic.
Fisher, W. P., Jr. (1991). Bettelheim's test. *Rasch Measurement Transactions, 5*(3), 164–165.
Fisher, W. P., Jr. (2002). Bettelheim's test revisited. *Rasch Measurement Transactions, 16*(3), 886–887.
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). Brookfield: Ashgate Publishing.
Fisher, W. P., Jr. (2003b). Mathematics, measurement, metaphor, metaphysics: Parts I and II. *Theory and 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. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement, 42*(9), 1278–1287.
Fisher, W. P., Jr. (2010a). Reducible or irreducible? Mathematical reasoning and the ontological method. In M. Garner, G. Engelhard Jr., W. P. Fisher Jr., & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 1, pp. 12–44). Maple Grove: JAM Press. [Reprinted from Fisher, W. P., Jr. (2010). *Journal of Applied Measurement, 11*(1), 38–59.].
Fisher, W. P., Jr. (2010b). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics Conference Series, 238*(1), 012016.
Fisher, W. P., Jr. (2012). What the world needs now: A bold plan for new standards. *Standards Engineering, 64*(3), 1 & 3–5.
Fisher, W. P., Jr. (2013). Imagining education tailored to assessment as, for, and of learning: theory, standards, and quality improvement. *Assessment and Learning, 2*, 6–22.
Fisher, W. P., Jr. (2017). A practical approach to modeling complex adaptive flows in psychology and social science. *Procedia Computer Science*, *114*, 165–174.
Fisher, W. P., Jr., & Cavanagh, R. (2016). Measurement as a medium for communication and social action, I & II. In Q. Zhang & H. H. Yang (Eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 153–182). Berlin: Springer.
Fisher, W. P., Jr., & Stenner, A. J. (2011a). Integrating qualitative and quantitative research approaches via the phenomenological method. *International Journal of Multiple Research Approaches, 5*(1), 89–103.
Fisher, W. P., Jr., & Stenner, A. J. (2011b). *Metrology for the social, behavioral, and economic sciences* (Social, Behavioral, and Economic Sciences White Paper Series). Washington, DC: National Science Foundation. Retrieved from http://www.nsf.gov/sbe/sbe_2020/submission_detail.-cfm?upld_id=36.
Fisher, W. P., Jr., & Stenner, A. J. (2016). Theory-based metrological traceability in education: a reading measurement network. *Measurement, 92*, 489–496.
Fisher, W. P., Jr., & Stenner, A. J. (2017). Ecologizing vs modernizing in measurement and metrology. *Journal of Physics Conference Series*, in press.
Fisher, W. P., Jr., & Wilson, M. (2015). Building a productive trading zone in educational assessment research and practice. *Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana, 52*(2), 55–78.
Fisher, W. P., Jr., & Wright, B. D. (1994). Applications of probabilistic conjoint measurement. *International Journal of Educational Research, 21*(6), 557–664.
Gadamer, H.-G. (1981). *Reason in the age of science* (T. McCarthy, Ed.) (F. G. Lawrence, Trans.) (Vol. 2, Studies in Contemporary German Social Thought.) Cambridge: MIT Press.
Gadamer, H.-G. (1989). *Truth and method* (J. Weinsheimer & D. G. Marshall, Trans.) (Rev. ed.). New York: Crossroad (Original work published 1960).
Gadamer, H.-G. (1991). *Plato's dialectical ethics: Phenomenological interpretations relating to the Philebus* (R. M. Wallace, Trans.). New Haven: Yale University Press.
Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago: University of Chicago Press.

Galison, P. (2008). Image of self. In L. Daston (Ed.), *Things that talk: Object lessons from art and science* (pp. 256–294). New York: Zone Books.
Gelven, M. (1984). Eros and projection: Plato and Heidegger. In R. W. Shahan & J. N. Mohanty (Eds.), *Thinking about Being: Aspects of Heidegger's thought* (pp. 125–136). Norman: Oklahoma University Press.
Golinski, J. (2012). Is it time to forget science? Reflections on singular science and its history. *Osiris, 27*(1), 19–36.
Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the next generation science standards for both formative and summative assessment* (K-12 Center at Educational Testing Service No. Invitational Research Symposium on Science Assessment). Princeton: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/gorin-mislevy.pdf.
Guttman, L. (1994). *Louis Guttman on theory and methodology: Selected writings* (S. Levy, Ed.). Dartmouth Benchmark Series. Brookfield: Dartmouth.
Hall, W. J., Wijsman, R. A., & Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Annals of Mathematical Statistics, 36*, 575–614.
Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14*(2), 75–96.
Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham: Kluwer.
Hambleton, R. K., Swaminathan, H., & Rogers, L. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Publications.
Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.
Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). New York: Harper & Row.
Heidegger, M. (1967). *What is a thing?* (W. B. Barton, Jr. & V. Deutsch, Trans.). South Bend: Regnery/Gateway.
Husserl, E. (1970). *The crisis of European sciences and transcendental phenomenology: An introduction to phenomenological philosophy* (D. Carr, Trans.). Evanston: Northwestern University Press.
Kant, I. (1970). *Metaphysical foundations of natural science* (J. Ellington, Trans.). Indianapolis: Bobbs-Merrill (Original work published 1786).
Kisiel, T. (1973). The mathematical and the hermeneutical: On Heidegger's notion of the apriori. In E. G. Ballard & C. E. Scott (Eds.), *Martin Heidegger: In Europe and America* (pp. 109–120). The Hague: Martinus Nijhoff.
Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. New York: Cambridge University Press.
Latour, B. (1993). *We have never been modern*. Cambridge: Harvard University Press.
Latour, B. (2004). How to talk about the body? The normative dimension of science studies. *Body and Society, 10*(2–3), 205–229.
Latour, B. (2005). *Reassembling the social: An introduction to Actor-Network-Theory.* (Clarendon Lectures in Management Studies). Oxford: Oxford University Press.
Lenoir, T. (Ed.). (1997). *Instituting science: The cultural production of scientific disciplines*. Stanford: Stanford University Press.
Levinsohn, F. H., & Wright, B. D. (1976). *School desegregation: Shadow and substance*. Chicago: University of Chicago Press.
Linacre, J. M. (1998). Ben Wright: The measure of the man. *Popular Measurement, 1*, 23–25.
Linacre, J. M. (2017). *A user's guide to WINSTEPS Rasch-Model (Version 4.00) [Computer software]*. Chicago: Winsteps.com.
Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 61*(4), 1–49.
Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin, 58*, 122–131.

Mari, L., Maul, A., Irribarra, D. T., & Wilson, M. (2016). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement, 100*, 115–121.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement, 51*, 315–327.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Masters, G. N. (1984). Constructing an item bank using partial credit scoring. *Journal of Educational Measurement, 21*(1), 19–32.

Masters, G. N. (1994). KIDMAP—a history. *Rasch Measurement Transactions, 8*(2), 366.

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

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology, 10*(5), 639–667.

Miller, P., & O'Leary, T. (2007). Mediating instruments and making markets: Capital budgeting, science and the economy. *Accounting, Organizations, and Society, 32*(7–8), 701–734.

Mundy, B. (1986). On the general theory of meaningful representation. *Synthese, 67*, 391–437.

National Research Council. (2006). *Systems for State Science Assessment.* Committee on Test Design for K-12 Science Achievement. M.R. Wilson and M.W. Bertenthal (Eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nersessian, N. J. (2002). Maxwell and "the method of physical analogy": Model-based reasoning, generic abstraction, and conceptual change. In D. Malament (Ed.), *Reading natural philosophy: Essays in the history and philosophy of science and mathematics* (pp. 129–166). Lasalle: Open Court.

Nersessian, N. J. (2006). Model-based reasoning in distributed cognitive systems. *Philosophy of Science, 73*, 699–709.

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

Nielsen, A. R. (1968). *Lust for learning*. Thy: New Experimental College Press.

Olitsky, S., Flohr, L. L., Gardner, J., & Billups, M. (2010). Coherence, contradiction, and the development of school science identities. *Journal of Research in Science Teaching, 47*(10), 1209–1228.

Overton, W. F. (2015). Processes, relations and relational-developmental-systems. In W. F. Overton & P. C. M. Molenaar (Eds.), *Theory and Method. Vol. 1 of the Handbook of child psychology and developmental science* (7th ed., pp. 9–62). Hoboken: Wiley.

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement, 71*, 46–55.

Petersen, A. (1968). *Quantum physics and the philosophical tradition*. Cambridge: MIT Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen: Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321–333). Berkeley: University of California Press.

Rasch, G. (1973/2011). All statistical models are wrong! Comments on a paper presented by Per Martin-Löf, at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark, May 7-12, 1973. *Rasch Measurement Transactions, 24*(4), 1309.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58–94.

Ricoeur, P. (1967). *Husserl: An analysis of his phenomenology* (J. Wild, Ed.) (E. G. Ballard & L. E. Embree, Trans.). *Northwestern University Studies in Phenomenology and Existential Philosophy*. Evanston: Northwestern University Press.

Ricoeur, P. (1970). *Freud and philosophy: An essay on interpretation*. Evanston: Northwestern University Press.

Ricoeur, P. (1974a). Ethics and culture. In D. Stewart & J. Bien (Eds.), *Political and social essays by Paul Ricoeur* (pp. 243–270). Athens: Ohio University Press.
Ricoeur, P. (1974b). The project of a social ethic. (D. Stewart, Trans.). In D. Stewart & J. Bien, (Eds.). *Political and social essays by Paul Ricoeur* (pp. 160–75). Athens: Ohio University Press.
Ricoeur, P. (1974c). Violence and language. In D. Stewart & J. Bien (Eds.), *Political and social essays by Paul Ricoeur* (pp. 88–101). Athens: Ohio University Press.
Ricoeur, P. (1976). *Interpretation theory: Discourse and the surplus of meaning*. Fort Worth: Texas Christian University Press.
Ricoeur, P. (1991a). Life in quest of narrative. In D. Wood (Ed.), *On Paul Ricoeur: Narrative and interpretation* (pp. 20–33). New York: Routledge.
Ricoeur, P. (1991b). Narrative identity. In D. Wood (Ed.), *On Paul Ricoeur: Narrative and interpretation* (pp. 188–199). New York: Routledge.
Ricoeur, P. (1992). *Oneself as another*. Chicago: University of Chicago Press.
Roche, J. (1998). *The mathematics of measurement: A critical history*. London: The Athlone Press.
Romer, T. A. (2013). Nature, education and things. *Studies in the Philosophy of Education, 32*, 641–652.
San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3 PL model. *Psychometrika, 80*(2), 450–467.
Scheler, M. (1973). *Selected philosophical essays*. Evanston: Northwestern University Press.
Sisson, J. H. (2016). The significance of critical incidents and voice to identity and agency. *Teachers and Teaching: Theory and Practice, 22*(6), 1–13.
Somers, M. R. (1994). The narrative constitution of identity: A relational and network approach. *Theory and Society, 23*(5), 605–649.
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.
Stocking, M. L. (1989). *Empirical estimation errors in item response theory as a function of test properties* (Educational Testing Service Research Report 89–05, ERIC Document ED395027), Princeton: Educational Testing Service.
Stone, M. H., Wright, B., & Stenner, A. J. (1999). Mapping variables. *Journal of Outcome Measurement, 3*(4), 308–322.
Thurstone, L. L. (1959). *The measurement of values*. Chicago: University of Chicago Press, Midway Reprint Series.
Townes, C. H., Merritt, F. R., & Wright, B. D. (1948). The pure rotational spectrum of ICL. *Physical Review, 73*, 1334–1337.
Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of Modern Item Response Theory*. New York: Springer Verlag.
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). New York: Springer.
Welten, R. (2016). Community from the perspective of life. *Analecta Hermeneutica, 8*, 130–148.
Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2014). Student reading growth illuminates the common core text-complexity standard. *The Elementary School Journal, 115*(2), 230–254.
Wilson, M. (Ed.). (2004). *National Society for the Study of Education Yearbooks. Vol. 103, Part II: Towards coherence between classroom assessment and accountability.* Chicago: University of Chicago Press.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah: Lawrence Erlbaum Associates.
Wilson, M. R. (2009). Measuring progressions: assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(August), 716–730.
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.

Woolley, A. W., & Fuchs, E. (2011). Collective intelligence in the organization of science. *Organization Science, 22*(5), 1359–1367.
Wright, B., & Tuska, S. (1965a). The effects of institution on changes in self-conception during teacher training and experience. *Proceedings of the 73rd Annual Convention of the American Psychological Association*, 1954, 299-300. *American Psychologist, 20*, 466, (abstract).
Wright, B., & Tuska, S. (1965b). The price of permissiveness. *The Elementary School Journal, 65*(4), 179–183.
Wright, B., & Tuska, S. (1967). The childhood romance theory of teacher development. *The School Review, 75*, 123–154.
Wright, B. D. (1954). Emotional factors shaping child care relationships. *Human Development Bulletin* (University of Chicago Committee on Human Development), pp. 28–34.
Wright, B. D. (1957). *A simple method for factor analyzing two-way data for structure*. Chicago: Social Research, Inc..
Wright, B. D. (1958). On behalf of a personal approach to learning. *The Elementary School Journal, 58*(7), 365–375.
Wright, B. D. (1959a). Identification and becoming a teacher. *The Elementary School Journal, 59*, 361–373.
Wright, B. D. (1959b). What price honors? *The Elementary School Journal, 59*, 436.
Wright, B. D. (1960). Should children teach? *The Elementary School Journal, 60*, 353–369.
Wright, B. D. (1961a). 'Goals' and 'Values' revaluated. *American Journal of Psychology, 74*, 310–312.
Wright, B. D. (1961b). Love and hate in the act of teaching. *The Elementary School Journal, 61*, 349–362.
Wright, B. D. (1968a). Introduction. In A. R. Nielsen (Ed.), *Lust for learning* (pp. 11–15). Thy: New Experimental College Press.
Wright, B. D. (1968b). The Sabbath Lecture: Love and order. In A. R. Nielsen (Ed.), *Lust for learning* (pp. 65–68). Thy: New Experimental College Press.
Wright, B. D. (1968c). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85–101). Princeton: Educational Testing Service.
Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97–116.
Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3*(1), 281–288.
Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), *Measurement and personality assessment*. Elsevier: North Holland.
Wright, B. D. (1988a). The efficacy of unconditional maximum likelihood bias correction: Comment on Jansen, Van den Wollenberg, and Wierda. *Applied Psychological Measurement, 12*(3), 315–318.
Wright, B. D. (1988b). Georg Rasch and measurement. *Rasch Measurement Transactions, 2*(3), 25–32.
Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45, 52.
Wright, B. D. (1998). Georg Rasch: The man behind the model. *Popular Measurement, 1*, 15–22.
Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know*. Hillsdale: Lawrence Erlbaum Associates.
Wright, B. D. (2005). Dedication: Memories from my life. In N. Bezruczko (Ed.), *Rasch measurement in health sciences* (pp. 6–17). Maple Grove: JAM Press.
Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21*(4), 331–345.
Wright, B. D., & Bettelheim, B. (1957). Professional identity and personal rewards in teaching. *The Elementary School Journal, 57*, 297–307.

Wright, B. D., & Douglas, G. A. (1975). *Best test design and self-tailored testing* (Tech. Rep. No. 19). Chicago: MESA Laboratory, Department of Education, University of Chicago. Retrieved from http://www.rasch.org/memo19.pdf. (Research Memorandum No. 19).
Wright, B. D., & Douglas, G. A. (1977). Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement, 37*, 47–60.
Wright, B. D., & Evitts, S. (1963). Multiple regression in the explanation of social structure. *The Journal of Social Psychology, 61*, 87–98.
Wright, B. D., & Gardner, B. (1960). Effect of color on black and white pictures. *Perceptual and Motor Skills, 11*, 301–304.
Wright, B. D., Loomis, E., & Meyer, L. (1963). Observational Q-sort differences between schizophrenic, retarded, and normal preschool boys. *Child Development, 34*, 169–185.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.
Wright, B. D., Mead, R. J., & Ludlow, L. H. (1980). *KIDMAP: person-by-item interaction mapping (MESA Memorandum #29)*. Chicago: MESA Press. Retrieved from http://www.rasch.org/memo29.pdf.
Wright, B. D., & Mok, M. (2000). Understanding Rasch measurement: Rasch models overview. *Journal of Applied Measurement, 1*(1), 83–106.
Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*(1), 23–48.
Wright, B. D., & Sherman, B. (1963). Teachers' self-awareness and their evaluation of childhood authority figures. *The School Review, 71*, 79–86.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.
Wright, B. D., & Tuska, S. A. (1968). From dream to life in the psychology of becoming a teacher. *The School Review, 76*(3), 253–293.
Wright, B. D., & Yonke, A. M. (1989). *American University Studies, Series V: Philosophy. Vol. 82: Hero, villain, saint: An adventure in the experience of individuality*. New York: Peter Lang.
Zaner, R. (1981). *The context of self: A phenomenological inquiry using medicine as a clue*. Athens: Ohio University Press.

Chapter 15
Ben Wright: Quotable and Quote-Provoking

Mark Wilson and William P. Fisher, Jr.

Abstract In this chapter we gather together quotable statements by and about Ben Wright.

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-3-319-67304-2_16

M. Wilson (✉) • W.P. Fisher, Jr.
Graduate School of Education, University of California, Berkeley, Berkeley, CA, USA
e-mail: markw@berkeley.edu; wfisher@berkeley.edu

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_15

15.1 Quotables

Wright on Rasch (from Andrich, 1995).

> "Georg was unwilling to take traditional cliches for granted. That intrigued me. His impassioned conviction that we are going to think for ourselves, that we are not going to just believe what anybody else says, that we are not going to just do things the way others have done them, but are going to figure things out for ourselves, and only do what makes sense to us, only do what we are able to make sense out of, that really appealed to me. That's the kind of person I am. Georg was a kindred spirit."

Wright on Rasch (from Andrich, 1995).

"He [Rasch] went right to the observation and modelled it. I liked that idea very much. It was clean and clear, fresh and new, sensible and uncluttered. I listened to him and I thought, 'This makes sense, in fact, better sense than anything I have heard so far.'"

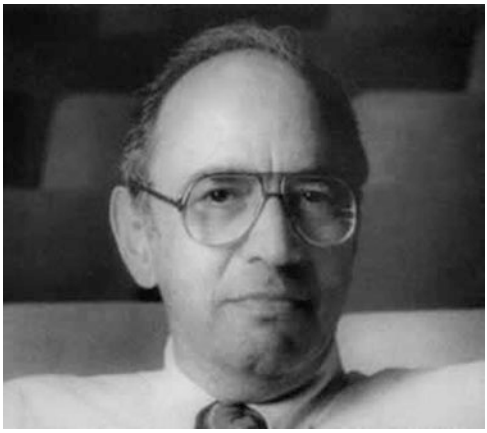
Wright recalling Rasch's 1960 visit to Chicago (from Andrich, 1995)

"So we made friends. After his lecture we had lunch at his CTS [Chicago Theological Seminary] apartment. He got out his cans of sardines, his brown bread, his pepper and his beer. He opened the sardines, put them on the bread, mashed them a little, poured on some oil and added lots of pepper. He enjoyed it all so much. He even enjoyed opening the can.

He was really into it. His pleasure in something as simple as a sardine sandwich was an inspiration to me. I thought, 'That's the way life should be. I like this man and the way he does things. I want to be like him.'"

Wright (1988) on the model-data relationship.

"As a young physicist in the 1940s, I did a lot of measuring… In physics, you keep collecting data until you get the data you want. You don't fit your theory or your ideas to the data that happens to be convenient…. You have demanding expectations about what you're doing. The aim is to find data to support your theory, not to find a theory that might fit all the data you might encounter."



Kuhn (1961/1977) on the model-data relationship

"…the scientist often seems … to be struggling with facts, trying to force them into conformity with a theory he does not doubt. Quantitative facts cease to seem simply 'the given.' They must be fought for and with, and in this fight the theory with which they are to be compared proves the most potent weapon."

Wright (1997) on measurement across fields.

"Today there is no methodological reason why social science cannot become as stable, as reproducible, and hence as useful as physics."

Mari and Wilson (2014) on measurement across fields:

"Rasch models belong to the same class that metrologists consider paradigmatic of measurement."

Luca Mari, Secretary of TC25 (Quantities and units),International Electrotechnical Commission (IEC)and member of the Joint Committee for Guides in Metrology

Wright (1997, p. 33) on measurement networks

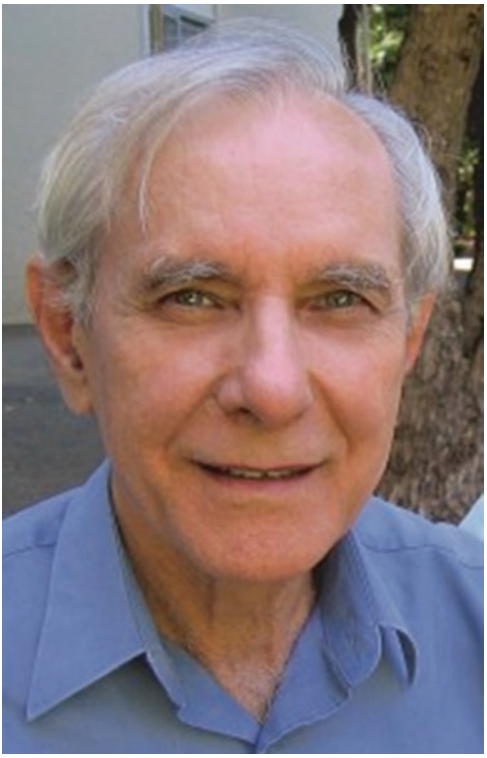
"Science is impossible without an evolving network of stable measures."

Latour (1987, p. 249) on measurement networks

"Every time you hear about a successful application of a science, look for the progressive extension of a network."

15.2 Quote-Provoking

David Andrich, University of Western Australia, Australia



If the Kuhnian shift from the statistical modeling paradigm to Rasch's measurement paradigm eventually "turns out to be a successful revolution, its *ultimate triumph* will have depended a great deal on the enthusiasm, energy, commitment and teaching of Ben Wright…"

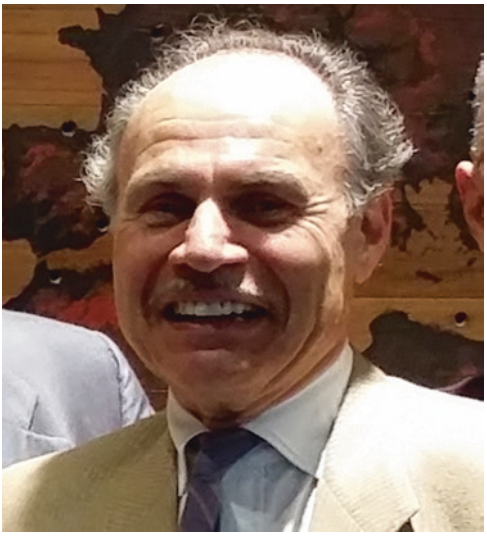
Betty Bergstrom, Pearson VUE, USA

"In the late 1980s I decided I wanted to explore Measurement as a career. In August, I called the University of Chicago to inquire and someone told me 'Call Ben Wright.' So I did (having no idea who Ben Wright was). Ben answered his phone and said, 'Well just come to my class, and if you like it, maybe you will stay.' I did come to his class, and I did like it and I stayed. And it changed my life."

Judy Beto, Dominican University. USA

"Ben was a great influence in my professional life pathways. He showed that qualitative data and methods are never so vivid as when they are in a measurement focus."

Nikolaus Bezruczko, Measurement & Evaluation Consulting, USA



"Ben was fundamentally successful in pointing the direction to better methods in social science research…. Objective science has lost a warrior, while those who worked with him have lost a loyal friend."

Trevor Bond, James Cook University, Australia

"Ben had answered repeatedly those same questions from battalions of beginners with indefatigable good humour and patience. The same beginner questions from sages who should know better often provoked his ire."

"…could Rasch measurement be where it is now, without his *passion*?"

"Ben had a disarming way of finding the Achilles' heel of an argument—amusing and salutary to watch, but devastating if you, personally, provided the object lesson."

Bill Boone, Miami University, USA

Ben's red pen comments on weekly memos:

- On checking data: Your scientific responsibility
- On hand analysis of data: Always a good idea
- Memo comments:
  - How very wonderful!
  - Take a stand! Be opinionated!
  - We invent in order to discover!
  - When we measure, we must choose an intention.
  - Onward and upward! (Royal, 2015)

George Engelhard, University of Georgia, USA

"I am still writing memos to Ben!"

Kathy Green, University of Denver, USA

"Ben allowed me to sit in his classes at the University of Chicago where I went on my first sabbatical leave—way back in the 1990s. We had a weekly meeting where he reviewed what I had written from that week, correcting my mistakes with a red pen, and generally being extremely kind to a novice. I use the examples and stories he told in class in my own Rasch model classes—'driving a Mercedes into the lake'[1]—an outright theft I think he would have approved of. He liked tangerines, so the price of weekly instruction was a tangerine. He shaped my interests and my career, and I thank him for that."

[1] Ben's point was that we should not allow data to be the sole criterion determining decisions on item quality and construct validity. If an item works well in its intended context, but breaks down when applied outside of that context, then perhaps it is being misused in a manner analogous to trying to employ a perfectly functional technology like an automobile in unintended and dysfunctional ways.

Ron Hambleton, Distinguished University Professor, University of Massachusetts, Amherst, USA

"I have often said that Ben, more than anyone else, inspired graduate students and faculty members, and specifically, inspired them to move the model and its applications forward."

"Our field is all the better for Professor Wright's impact. Today, Professor Wright's contributions can be found around the world and indeed, he was responsible for a paradigm shift in the advancement of measurement. That's something only a very few can claim. Professor Wright was a giant in the measurement field, and his contributions will be long remembered and valued." (Royal, 2015)

Roberta Henderson, Rosalind Franklin University of Medicine & Science, USA

Fond remembrances of days in Judd Hall:

- Yardstick interrogations
- One idea memo and one critique memo every week
- Red ink on assignments: more was better
- Gatherings at Ben's home
- Derive the Rasch model: "Now"
- Discussions of what is real
- Delight in return of former students from all corners of the world
- Anticipation and excitement over a new data set
- Enduring support (Royal, 2015)

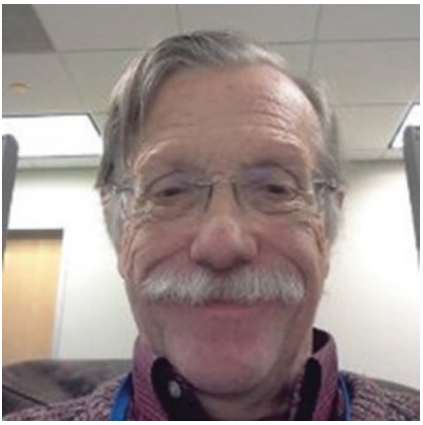
Svend Kreiner, University of Copenhagen, Denmark

"It is safe to say that without Ben Wright we would not be here today to celebrate the technical achievements of an obscure Danish mathematician."

Opening remarks at the Conference Celebrating 50 years since the publication of Rasch's *Probabilistic Models,* University of Copenhagen School of Business, 2010.

Mary Lunz and John Stahl, Pearson VUE & American Society for Clinical Pathology, USA

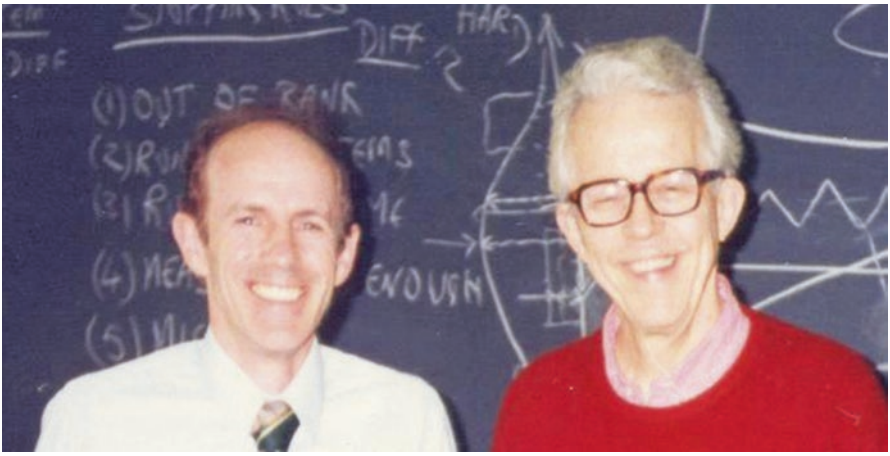
"We all applaud Ben for his dedication to the field of objective measurement. In fact many of us owe the success of our careers, in large part to Ben Wright, his research, his work…"



[Ben was] "a complex individual with many assets and liabilities. …tact was not one of his greatest assets."

Mike Linacre, University of the Sunshine Coast, Australia

[Teaching measurement in action in the classroom] "was where Ben really excelled. He would scrutinize the hierarchy of item difficulties. After some discussion with the student, Ben would have a definition of the latent variable the test or survey was actually measuring…."



…"Then Ben would investigate the misfitting items and persons. For the students, it was like watching a combination of a detective and a psychoanalyst working together. Why had this seemingly mundane MCQ item provoked some smarter students to respond the way they had? … Ben could discern the mental processes that produced even the foggiest data." (Royal, 2015)

Trudy Mallinson, George Washington University, USA

"Ben loved yardsticks. I found one that has only two markings on one side of it, at ¼ and ¾ portions of a yard. This unique, and somewhat puzzling, ruler reminded me of Ben's attentiveness to measuring devices and how, above all else, they should be useful. Yes, they should be accurate and consistent but the amount of precision represented by the device should be practical. You don't need a 36-inch ruler if the only things you measure are less than 12 inches long. And you don't need a ruler marked off in 288 1/8-inch units if the only things you need to measure come in lengths of ¼ and ¾ yards!" (Royal, 2015)

Ron Manheimer, The Manheimer Group (and fellow teacher with Ben at the New Experimental College in Denmark, 1967)

[While at the NEC] "Ben had a course called 'The Psychology of Being a Teacher'. It was a course for teachers. And it asked them, 'Why did you choose this career? What happened in your life?' And how it turned out was that people would either talk about when they were kids and they had this inspiring teacher and they wanted to be like them, or they had a terrible experience as a student, and now they wanted to fix it, by being a teacher and repairing this terrible experience they had themselves. The stuff that would come out was really interesting. It was helpful to free up the motivation. Most people had not thought about it in a long time, or never really thought about what had really happened to them, or lost touch with it. Ben was very smart. He had a sixth sense as a psychologist. He was astute. He could pick up on very subtle things about people. It was a gift." (Jakobsen, 2014)

Geoff Masters, Australian Council for Educational Research, Australia

"One morning in Chicago I arrived at my desk to find a note left by Ben Wright. In handwriting that filled most of the page Ben had written,

"G. Isn't science wonderful? B".

"I don't remember now what excited Ben that morning. He often took home what we were working on and brought it back next morning covered with ideas. Almost forty years later I still have that note—a reminder of the daily exhilaration of working with Ben as we pored over analyses of data sets, worked on the mathematics of measurement, and experimented with more succinct ways of expressing and explaining our work."

"I learnt a great deal from Ben. He gave me ways of thinking and writing and a passion for discovery that have stayed with me through my career. In a general sense, what Ben and I were attempting to do was to construct deeper meaning from the specifics of experience."

Bob Mislevy, Frederic M. Lord Chair in Measurement and Statistics, Educational Testing Service, USA

"I took several of Ben's Friday seminar classes. Hearing him think and discuss on the fly was very instructive—one of those things where you see it isn't all so cut and dry like stat books make it seem, but a constructive, active interplay between what's in the books and what's in the world, and how your philosophies and models are the bridge."

"Although Ben and Darrell [Bock] clearly thought differently about IRT, there was a lot one could learn from both of them. They were both on my dissertation committee. I worked closely with both of them, on different chapters. They were both supportive to me and cordial to each other throughout the process, which I didn't understand well enough at the time to appreciate as much as I should have!"

Magdalena Mok, Education University of Hong Kong

This ignorant beginner in Rasch modelling knocked on Ben Wright's door one day in 1997 with hundreds of questions about the model, half-expecting to be given a cold shoulder, as most famous-but-too-busy academics would, but instead, he greeted me warmly and said, 'You are most welcome to join my postgraduate class and come to my house in the afternoons for private tuition.' 'How much should I pay you for the private tutorial?' I murmured. After all, I was just a stranger to him 5 min ago. He then told me the story about his encounter with his teacher, George Rasch, and how they shared lunch—sardines, bread and orange juice—every afternoon in their journey of knowledge discovery.

He ended his story by saying, 'Just bring sardines and bread for two people. I will provide the orange juice for us.' Ben cherished those precious moments with his teacher as I relish the generosity of Ben to this day. Mother Teresa said, 'Let no one ever come to you without leaving better and happier.' Ben most definitely lived by those words.

Mark Moulton, Educational Data Systems, Inc., USA

"Ben was a man of vivid faults and even more vivid virtues, a great psychometrician and a greater friend." (Royal, 2015)

Carol Myford , University of Illinois at Chicago, USA

"I came to the University of Chicago with a strong interest in assessment in the arts. In Ben, I found a kindred spirit. Throughout my years of study with him, he encouraged me to passionately pursue those interests, even though the constructs I wanted to study were ones that were not easily defined or measured. He taught me not to shy away from the challenge of working with those elusive constructs. For those life lessons, I am eternally grateful."

Leslie Pendrill, SP Technical Research Institute of Sweden, Borås; past Chair, European Association of National Metrology Institutes

"The Rasch approach…is not simply a mathematical or statistical approach, but instead [is] a specifically metrological approach to human-based measurement." (Pendrill, 2014)

Georg Rasch (1972/1988) on Wright

"…the cooperation with Dr. B.W. … has given much inspiration to and been of great practical use for GR."

"…since his first visit to Denmark in 1964 BW has practiced an almost unbelievable activity in this field, and results have certainly not been lacking."

Matthew Schulz, Smarter Balanced Assessment Consortium, USA

"Ben Wright's approach to educational measurement recognizes the contribution that people who are not highly trained in mathematics and statistics can make, and want to make, through measurement, to a particular topic or discipline. The ideas Ben promoted through Rasch models are very simple and powerful. They do not require the practitioner, or even the statistician, to become preoccupied with the statistical details of psychometrics. Rather, they require the statistician and substantive expert alike to attend to a measure's internal, substantive meaning."

Jack Stenner, MetaMetrics, Inc., USA

Topics of Tuesday conversations with Ben through the 1990s

- The awesome power of the Gibbs/Einstein ensemble interpretation
- Causal vs. Descriptive Rasch Models
- How to compute fit statistics that are sample independent
- The Fahrenheit method for establishing a unit of measure
- Employing multiple measurement mechanisms to establish the reality (i.e. existence) of an attribute
- Using the trade-off property to test for quantitative status of an attribute (Royal, 2015)

Donna Surges Tatum, Meaningful Measurement, Inc., USA

"Ben Wright had a tremendous influence on my life. He helped me take a left turn from being a Rhetorician and transformed me into a Psychometrician. So now I tell stories with numbers as well as words."

"Ben acculturated his students with a collegial spirit and a collaborative approach to the science of measurement. This has served us well as we spread Rasch measurement to all disciplines all over the world."

Herb Walberg, University Scholar, University of Illinois, Chicago, USA

"My experience in the 55 years since I first met Ben lead me to believe that we are now beginning the 'Golden Age of Measurement' in education, psychology, and the social sciences. Decades ago, visionaries Georg Rasch, Ben Wright, Bruce Choppin, and others showed us the way."

"Can we measure up to their standards?"

Mark Wilson, Graduate School of Education, University of California, Berkeley, USA

"So now you can see that when I said Ben was my 'best reader,' I meant a lot more than perhaps was obvious. I meant that he had startled me, and inspired me, with other peoples' ideas, and, of course, with his own. I meant that he had shown me what it meant to take yourself and your ideas seriously, and had encouraged me to do just that. I meant that he had fostered my doubts, and lived out his own. I meant that he had shown that the academic life is definitely worth living, but you probably have to escape the academic world in order to fully enjoy it. Thank you Ben."

References

Andrich, D. (1995). Rasch and Wright: The early years (transcript of a 1981 interview with Ben Wright). In J. M. Linacre (Ed.), *Rasch measurement transactions, Part 1* (pp. 1–4). Chicago: MESA Press. Retrieved from http://www.rasch.org/rmt/rmt0.htm.

Jakobsen, J. (2014). Interview with Ron Manheimer. In J. Jakobsen (Ed.), *New Experimental College Tabloid* (pp. 57–62). Aarhus: Kunsthal Aarhus. Retrieved from http://kunsthalaarhus.dk/en/jakob-jakobsen.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis, 52*(168), 161–193. (Rpt. in T. S. Kuhn, (Ed.). (1977). *The essential tension: Selected studies in scientific tradition and change* (pp. 178–224). Chicago: University of Chicago Press.).

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. New York: Harvard University Press.
Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement, 51*, 315–327.
Pendrill, L. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science, 9*(4), 22–33.
Rasch, G. (1972). Review of the cooperation of Professor B. D. Wright, University of Chicago, and Professor G. Rasch, University of Copenhagen; letter of June 18, 1972. *Rasch Measurement Transactions, 2*(2), 19. Retrieved from http://www.rasch.org/rmt/rmt22c.htm.
Royal, K. (2015). A tribute to Benjamin D. Wright [Special issue]. *Rasch Measurement Transactions, 29*(3), 1528–1546. Retrieved from http://www.rasch.org/rmt/rmt293.pdf.
Wright, B. D. (1988). Georg Rasch and measurement. *Rasch Measurement Transactions, 2*(3), 25–32. Retrieved from http://www.rasch.org/rmt/rmt23a.htm.
Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45. 52. Retrieved from http://www.rasch.org/memo62.htm.

Erratum to: Psychological and Social Measurement: The Career and Contributions of Benjamin D. Wright

Mark Wilson and William P. Fisher, Jr.

Erratum to:
M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*, Springer Series in Measurement Science and Technology, https://doi.org/10.1007/978-3-319-67304-2

The original version of this book included erroneous affiliations that were inadvertently introduced during the production process. These incorrect affiliations appeared in the Front matter and in Chapters 1, 8, 14 and 15 of the original version.

The correct affiliations are shown below and the book has now been updated to reflect this correction:

Editors
Mark Wilson
Graduate School of Education
University of California, Berkeley
Berkeley, CA, USA

William P. Fisher, Jr.
Graduate School of Education
University of California, Berkeley
Berkeley, CA, USA

The updated online version of these chapters can be found at
https://doi.org/10.1007/978-3-319-67304-2_1
https://doi.org/10.1007/978-3-319-67304-2_8
https://doi.org/10.1007/978-3-319-67304-2_14
https://doi.org/10.1007/978-3-319-67304-2_15

The updated online version of this book can be found at
https://doi.org/10.1007/978-3-319-67304-2

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2_16

Appendix A: Love and Order—A Sabbath Lecture

Benjamin D. Wright

Abstract This is the full text of Wright's Sabbath Lecture, delivered 16 September 1967 at the New Experimental College in Thy, Denmark. Reprinted from Nielsen, A. R. (1968). *Lust for learning.* Thy, Denmark: New Experimental College Press, pp. 65–68.

This is OUR *Sabbath Lecture* here at New Experimental College. I have previously experienced two Sabbath Lectures which have taught me something of what *Sabbath Lecture* means. They have taught me that I must take the words *lecture* and *sabbath* most seriously—and in their full and demanding meaning.

Lecture means that I must step forward myself, alone, and take my own full responsibility to create and compose a talk to give to you. In this *lecture* I cannot permit myself to take that other easier, more amusing *discussion* way of titillating a fresh slice of our relationship with each other and then leaning on our common wit and lust for life to make something exciting out of it. No! In this *lecture* I must take all the responsibility for making something worthwhile out of this meeting—and all the responsibility for wasting your time.

I know I cannot do it! I know that I cannot succeed—that some of your time will seem wasted. John Littlewood tells me that *Sabbath Lectures* are boring, and that nothing can be done about it. If he is right and they are bound to be boring, then I must face my responsibility to do a good job boring you.

Sabbath? What does that mean? A holy celebration of our life together. That time of the week when we rest from our various labors and loves and gather together to reflect on what we have done, what we are doing, and on what we have in common.

So my text must bear on our life together. It must truly be of our celebration. It cannot be a casual report of just my concerns, of just my research or even of just my troubles. My text must bear on what we have in common, on our past and present together, on our research and even on our troubles. But where will I find that—where

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2

are our common problems? Somewhere between us out there in the air of the room? Somewhere unseen by me in you? Somewhere where I cannot know them?

No, our problems that I must celebrate in my text can only exist for me in myself. I am forced back on myself, and must find our problems there. But now my perspective is different from that of the casual report that I already rejected. Now I know I must find the consequences of our life together, of you, in me in my work and bring these consequences back to you in my Sabbath Lecture.

I must allow what you have done to me to work inside, to disturb and revise what I have been doing. I must discover how you live in my work, in my research and in my troubles. I must find a way to translate our life together into the fixed words of my work and my work into the flowing feelings of our life. And even if I cannot do it, I must preach about it. Well, I know I am not ready to succeed in that—I cannot do it—but I want to work on it and pray that you will see some ways to help me with it.

My text is the conflict of *love* and *order*. What are they? To bring this out I will oppose them and through their contrast try to clarify their distinctive characters.

If love is feeling, then thought is order. If order is action and structure, then sensation and flow belong to love. In human relating we think of responsiveness and reunion as loving. Then self-assertion and individuation are orderly. Communion is an expression of love. Identity is an assertion of order. Order is the foundation of clarity. But love can lead to confusion. Love nourishes hope and creation, and promotes the rich experience of life. But order aims at discipline and conclusion and requires in the end the fixed settlement of death.

Love and order also play their part in the practice of science. Full experience of reality is an expression of love. But that narrowing and selection of experience which becomes scientific observation is an act of order. The control and organization of observations which become theory are triumphs of order. But the response to theory which becomes understanding and insight are celebrations of love.

In perception order is an achievement of taste, touch and sight. We taste the difference, get hold of the issues and see our way clear. Discrimination in taste, a firm grip on reality and clear vision are the perceptual acts which make orderly thought possible. But it is love which is prominent in the diffuse and involuntary aspects of smell and hearing. Smelling and hearing cannot be well focused. We cannot turn them on and off at will. We are overwhelmed by an odor and easily seduced by sweet music or forked tongues.

So what are love and order? Order can make room for love, can trans-form aggression into the lord protector of a happy kingdom. But order also contains the seeds of death and the institutionalization of anger into hate. Love in contrast is a spontaneous expression of feeling, including anger. Love is not so much an act as an ever changing experience. We are by order arranged but by love possessed.

How might love and order be used in the management and enjoyment of human relations? To pursue this I will talk about the *theory* of order and the *practice* of love. The theory of order says that all acts are adaptive creations of the administrating ego. They are over determined to meet, as they do, as many needs as possible.

The practice of love then is to view any act as an accomplishment, to ask what are its benefits to the actor and to expect more than one benefit.

The theory of order says that *negation* is the equivalent of unconscious repression. So the practice of love is to deal with a block in communication due to repression by asking what is *not* relevant, what is *not* the case.

Theory says that *projection* is a general and basic mode of inter-personal perception. Practice then is to know that assertions about others are self-descriptions and to be ready to discuss, formulate and clarify re-pressed issues in terms of the others upon whom we project.

Theory says that in *transference* we treat new experiences as though they were repetitions of old ones, that we re-enact old important relations with new others. Practice is then to recognize in today's relations yester-day's problems.

Finally the theory of order says that we master overwhelming and frightening experiences by identifying with them in a reversal of voice called identification with the aggressor. The practice of love leads us to ask *who treated us the way we're treating others?*—to recognize old aching wounds in new attacks.

When I contrasted *lecture* and *discussion*, I was feeling the conflict of *love* and *order*. The good discussion is a kind of love play where lyric spontaneity and playful discovery reign and order is kept at rest in the wings. But the lecture is something different. In the lecture it is the epic reign of order which prevails. Love no longer plays on center stage. Passion appears only in perspective. Love and emotion are recollected in tranquility.

When I was forced to realize that a Sabbath celebration meant neither exhibition of myself nor play with you, but rather thoughtful communion together—then I began to discover what is the communion of love and order.

It is order reworked for the sake of love, and love harvested for the nourishment of order—order for love *and* love for order. If this communion has a sacrament, then it will be found in the fruitful human relation.

Appendix B: Should Children Teach?

Benjamin Wright

Abstract This article provides an overview of the history of peer learning, which has recently become a matter of renewed interest among educators. Also of special interest in this article are Wright's comments on mechanical teaching aids. What he has to say resonates clearly with contemporary concerns with the roles of computer technology in learning. Just as Wright's personal approach to learning and his dissatisfaction with the educational measurement techniques of the day presaged his response to Rasch, so, too, do the issues concerning peer learning and mechanical learning aids in this article foreshadow Wright's enthusiasm for Ben Bloom's "two-sigma problem" (as was recounted by Sophie Bloom in her preface to Guskey's collection of Bloom's students' memories published by Rowman & Littlefield in 2006). The following article is reprinted in full with permission from the publisher, the University of Chicago Press:Wright, B. D. (1960). News and comment: Should children teach? *The Elementary School Journal 60*(7), 353–369.

In education it was the best of times; it was the worst of times. School for everyone was in the air. New methods were in ferment. But a population explosion threatened to swamp, if not to sweep away, the schools. There was a paralyzing teacher shortage. Education was ripe for revolution.

The time? One hundred and fifty years ago. The revolution? A method of instruction called the monitorial system. The essence of the idea was to let the children help the teacher by teaching one another.

At first the monitorial system flourished. It played a vital role in the birth of public education and the organized preparation of teachers. Later under the scorn of a new generation of reformers the plan withered and vanished as an ordained method.

Since the passing of the early nineteenth-century champions, few accounts of the monitorial system have been free of disdain. Yet a careful examination of early descriptions of the system—not to mention an honest appraisal of one's own teaching and learning experience—suggest; that mutual instruction is an idea of great educational power.

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2

The Voice of Experience

Informal mutual instruction is surely as old as society. As early as the first century the great Roman teacher, Quintilian, pointed out in his *lnstitutio Oratoria* how much the younger children can learn from the older children in the same class. In Hindu schools the use of mutual instruction dates back to ancient times.

Formal recommendations to use mutual instruction began to appear at the end of the Renaissance. In the 1530s, Valentin Trotzendorf, a German teacher, used his more advanced pupils in the government and the instruction of the others in his school at Goldberg in Silesia. When the Spanish Jesuits opened the College of Lisbon in 1553, they organized a system of "decurions," in which each group of ten pupils was led by a student monitor. By 1591 the decurion system was a formal part of the *Ratio Studiorum,* the Jesuit code of liberal education.

A few years later the English schoolmaster', John Brinsley, in his book *The Grammar Schoole,* which appeared in 1611, described his use of "two or foure Seniors in each fourme … for overseeing, directing, examining, and fitting the rest [of the children] in every way." In the 1630s the great Moravian teacher, John Comenius, observed in his *Didactica Magna:*

> The saying, "He who teaches others, teaches himself," is very true, not only because constant repetition impresses a fact indelibly on the mind, but because the process of teaching in itself gives a deeper insight into the subject taught … The gifted Joachim Fortius used to say that … if a student wished to make progress, he should arrange to give lessons daily in the subjects which he was studying, even if he had to hire his pupils John Amos Comenius, *The Great Didactic,* translated by M. W. Keatinge. London: Adam and Charles Black, 1896].

As a result, when Comenius discussed "How can a single teacher teach a number of boys … at one time?" he recommended as both necessary and educationally preeminent a regular system of mutual instruction.

This method of helping the teacher was also used with very young children before the eighteenth century. Jean Baptiste de la Salle, who founded the Christian Brothers to educate young children, outlines in his *Conduite des Ecoles* the monitorial system he used at Rheims in the 1680s. The Reverend john Barnard said in his autobiography that it happened to him when he was a 5-year-old schoolboy in Massachusetts in 1686. But it was not until the late eighteenth century, when the Industrial Revolution spawned intense public interest in education, that mutual instruction became widely publicized.

In 1791, the Anglican cleric Andrew Bell took charge of a boys' orphanage in Madras, India. Bell found himself unable to influence the adult teachers available to teach his children properly. Having observed the Hindu system of mutual instruction, he turned to his boys for help and discovered that they could be excellent teachers to one another.

Bell recorded his experience in *An Experiment in Education,* published in 1797, and considered himself the inventor of monitorial instruction. But the man who most vehemently and successfully claimed the "new" idea for his own and

who did the most to spread it as a revolution in education was Joseph Lancaster, an English Quaker.

In 1798 Lancaster opened a school for poor children in London. He intended to hire adult assistants to help him teach but could not raise the money. As a result he was forced to see whether the children themselves could help one another. Like Bell, Lancaster was so overwhelmed with the constructive consequences of this invention that in 1803 he wrote a book, *Improvements in Education,* describing his experiences and devoted the rest of his life to telling the world about the new educational method.

Lancaster lectured passionately/on the monitorial system in Britain, the United States, and South America. His personal endeavors were beset by difficulties, since his projects invariably exceeded his resources, but the ideas he spread and the schools he caused to be established were impressively popular for some 30 years.

Unfortunately, the method of using pupils to teach one another invites an economic hallucination. Unlike Bell, who turned to the children because he was unable to persuade his teachers to teach the way he wanted them to, Lancaster was forced into the monitorial system because he could not afford to hire teachers.

He soon found that the potential economies of the method made it extraordinarily appealing to the wealthy and the governing. Instead of pursuing and developing the educational potentialities of the system, he was seduced by his desire to win his audiences. As a result, he concentrated unduly on the economic advantages of monitorial instruction.

Soon he was envisioning schoolrooms of a thousand pupils guided unerringly by a single overseeing schoolmaster and a complex military hierarchy of monitors. The vision was sensational and probably played the leading role in the wildfire spread of the Lancasterian System. But it also led to the eventual demise of organized mutual instruction in the nineteenth century.

The wealthy were happy to offer education to the poor on such painlessly inexpensive terms, but the poor became increasingly unhappy about getting for their children what they viewed as second-rate.

In the pursuit of economy, Lancaster grossly overmechanized his system. The educational idealists of the day objected and argued that good adult teachers were better than children any time.

These protests were somewhat irrelevant at first since there were hardly any teachers-good or bad. By the mid-nineteenth century, however, the growing supply of teachers and the combined pressures of organized labor, the consciences of the rich, and the ideals of the pure in mind led to the birth of public education and the end of the monitorial system.

Economy is not the essential virtue of mutual instruction. We should not be deluded into concluding that the failure of the economy-sized monitorial system in the nineteenth century means that we should avoid mutual instruction in the twentieth.

An Amazing American

Some nineteenth-century proponents of mutual instruction were not deluded by its unhappily treacherous economic appeal. William Bentley Fowle was one of these. His lecture on the monitorial system given at Andover in 1846 appears in his book, *The Teacher's Institute* (New York: A. S. Barnes, 1875, pp. 185–207). We can turn to Fowle for a vivid and penetrating account of what mutual instruction was in his school and what it could be in ours.

Fowle grew up in Boston. His first experiment with the monitorial system dates back to the early 1820s when he found himself with a school for uneducated poor children on his hands and no teacher. Fowle, who was then a printer and bookseller, took the teacher's place temporarily rather than deprive the children of school. But since no other teacher could be found, Fowle ended up serving for several years as schoolmaster to well over a hundred boys and girls of all ages.

Fowle's work was so impressive that in 1827 a group of Bostonians sought him to organize a girls' private school along the same lines. This school of about a hundred pupils he taught on his own from 1827 until 1840.

The school and its master were remarkable in many ways. Fowle was one of the first in this country to emphasize Pestalozzi's natural method of teaching. The school had a library of over six hundred volumes and more than a thousand dollars worth of apparatus with which to teach science and do laboratory experiments. Fowle was also one of the first to introduce blackboards and daily physical exercise and to abolish corporal punishment. Of other schools more typical of his day he said:

The best disciplined minds are often found in those children, who, by what the world terms a misfortune, are thrown upon their own resources, and early accustomed to the exercise of their moral and intellectual faculties …. Do I err when I say that no good opportunity for such exercise is afforded in common schools, where each is required to hoard up knowledge, and is forbidden to impart it to others, where intercourse is prohibited, and whispering is high treason, where change of place, if not of position, is punished as depravity, where implicit obedience is the divine right of the teacher, and the divine wrong of the pupil; where, in fact, the best pupil is he who most nearly resembles an automaton?

On the values of students' teaching Fowle says:

> Teaching is learning, and learning of the very best kind. I appeal to teachers and ask, whether every faithful attempt to teach the children under their care does not increase and improve their own knowledge.
>
> By teaching the younger children, the more advanced are constantly reviewing their studies, not by learning merely, but by the surer method of teaching what they- have learned to others.

Turning to the qualifications of a child to teach, Fowle says:

> If a child may not teach what he does know to one who knows less, because his knowledge is limited, I do not see but all teaching must cease …. The wisest and best of us go to church, and to lectures, without repugnance, although we know that the preacher or the lecturer is only a monitor, who knows, perhaps, a little more than we do of the subject under consideration, but who would perhaps come to us for information on many other subjects.

> The art of teaching depends more upon adapting the explanation to the capacity of the learner than upon the amount of knowledge accumulated by the teacher. Is it unreasonable then to suppose that the explanations of children may sometimes be better suited to the understanding of children than those of adults would be? I am not ashamed to own that I often called on my monitors to explain what I had failed to make a little scholar apprehend.

Still another value of the system for children, according to Fowle, is its natural give and take:

> The monitors in every branch are the best pupils in that particular branch and every monitor may also be a pupil of his fellow-scholar, as he is of the master …. One hour, he may govern his class according to fixed laws enacted by the master, and well understood by every pupil; the very next hour, he may be subject to one of the very pupils that he had just directed. The monitorial plan, as I used it, is the true democratic one: the children all had a chance at the offices, though only the qualified and the deserving were appointed. Being sometimes governed, children are less likely to become imperious; and sometimes commanding, they will not too easily become servile.

Fowle appraised his pupils' skills carefully and appointed only the proficient as pupil-teachers or monitors. But he managed to give nearly every child a chance:

> The ingenious teacher will, at times, make monitors of all his pupils … I often employed my second class in showing beginners how to study their lessons; a duty that teachers themselves are too apt to neglect …. No child, but the very lowest, was so low that she could not teach something, and that something I always required her to teach.

For anyone who has lived with children the educational benefits of mutual instruction are apparent. Fowle found that it was the rate child who could not teach something to his classmates. Do we have in mutual instruction an obvious and promising way to alleviate the present shortage of teachers?

Today we hear a great deal about the neglected talents of the gifted. We are concerned about our gifted children as resources for tomorrow. Why wait? Why limit the present richness of their gifts? Why deprive these children of the solid wisdom and warm satisfaction that could grow from their working as assistant teachers today?

I have spoken of two promises of mutual instruction: its promise for helping children learn and its promise for relieving today's teacher shortage. I want to devote the rest of this editorial to a third promise.

From the monitorial schools of the early nineteenth century came the first normal schools-the first organized teacher training institutions. Mutual instruction has great power not merely for relieving the classroom problem of today but also for solving the career problem of teachers for tomorrow.

Classrooms and Careers

Modem proposals for solving the teacher shortage tend to fall into two groups: those that concentrate on classrooms and those that concentrate on careers.

"We do not have enough teachers. How can we best use available resources to educate our children?" This is the question asked by those whose first concern is with today's classrooms.

"Not enough young people are becoming teachers. How can we create more teachers for tomorrow?" This is the question posed by those who are concerned chiefly with the question of careers.

Why this divided attack on the teacher shortage? The reasons are clear enough. Overflowing classrooms are an explicit crisis. They are the problem our schools face today. Careers are more intangible. They are faced in college. They belong to tomorrow.

This is a treacherous division. Immediate emergencies may force us to concentrate on manning classrooms now. But we must not allow the urgencies of the moment to cloud our better judgment. If crowded classrooms are the symptom, a shortage of young people interested in teaching as a career is the cause.

Most proposals acknowledge that unless we prepare more teachers, we will one day be in serious trouble. But few proposals undertake the next step-consideration of how experience in today's classrooms affects tomorrow's careers. What connection is there between the creation of teachers for tomorrow and the way we keep our schools today?

Teachers and teachers-to-be can help us answer this question.

When do young people decide to become teachers? Why?

When Is Teaching Chosen?

The incentives we now rely on to attract young people to the teaching profession are directed toward the late high school and college years. We hear of better vocational guidance in the last years of high school, better scholarship and loan programs in college, more appealing and more effective teacher-preparation programs, higher salaries, and better working conditions.

These particular incentives imply that the decisive moment for choosing teaching comes late-in college. But what if the decision comes earlier? What if important decisions about teaching are made in high school or even in elementary school? Then the appeals we now rely on may miss the mark. They may come too late. If the decision to teach is born early, we will want to offer incentives before college, perhaps even before high school.

In the 1920s, Charles Valentine asked 348 British teachers in training when they chose their profession. In the November, 1934, *British Journal of Educational Psychology* he reported that 56% said between the ages of 15 and 17, and 90% before the age of 20.

This finding is not peculiar to the British Isles. Authors of the Yale-Fairfield Study *Report for 1954–1955* state that, among a national sample of 1,066 college Seniors preparing for elementary-school teaching, 68% reported that their decision was made before leaving high school. Clarence Fielstra reported in the May, 19; 5, *Journal of Educational Research* that, among 230 California Juniors and

Seniors in an introductory course in education, exactly *50%* made the decision before graduation from high school.

The lowest figure I could find was reported by Robert Richey and William Fox in the May, 1948, *Bulletin of the School of Education, Indiana University.* Among a group of one hundred Indiana Freshmen who definitely intended to become teachers, only 37% said that they made the decision before leaving high school.

This last piece of evidence, however, must be qualified. In this study the time of decision was not obtained by a direct question but was inferred by Richey and Fox from answers to other questions. Thus 37% represents a rock bottom estimate.

What about the decisions made even earlier? R. H. Morrison and S. D. Winans, in their 1949 monograph for the New Jersey State Department of Education, *Choosing Teaching as a Career,* report that 40% of a group of 1,423 applicants to New Jersey teachers colleges said that they had made their decision to become a teacher before the tenth grade. Among the Yale-Fairfield national sample of students, 24% reported that they made their decision before leaving elementary school. Isobel Willcox and Hugo Beigel said in the June, 1953, *Journal of Teacher Education* that 21% of 152 New York Freshmen in teacher education traced their desire to teach back to childhood.

So much for early decisions among undergraduate students in education. What about high-school students? Richey and Fox wrote in the July, 19 51, *Bulletin of the School of Education, Indiana University,* that of 970 Indiana high-school students who had "considered" teaching, 48% had done so before high school; of 261 who had "decided" on teaching, 35% had done so before high school.

Among experienced teachers, the evidence that career decisions are made early is even stronger. Lawrence Stewart asked a group of 260 summer graduate students, *50%* of whom had been teaching in the South for five years or more, when they decided to become teachers. He reports in the January, 1956, *Peabody Journal of Education* that 25% answered "before high school," 66% "before college." Among a group of 839 teachers, most of them from the North, questioned in the Yale-Fairfield Study, 30% said they decided before high school and 70% before college.

By now we may be ready to agree with Richey and Fox when they urge that encouragement "to give consideration to the selection of teaching as a vocation … should start very early in the school life of the child and should continue as long as he remains in the public schools" (1948). It seems pretty conclusive that among those who do make a ~decision to become teachers, 20–30% make their decision in elementary school, 40–70% have made their decision before they leave high school.

But what about those who decide not to become teachers? When do they make their decision? When are they lost to the teaching profession?

Roderick Langston and William Nutting asked 3,140 Oregon school children what they thought of elementary- school teaching as a vocation. The *Journal of Teacher Education* for June, 19 51, reported the results. Sixth-graders, 908 strong, said elementary- school teaching looked attractive to them, but 818 ninth-graders and 581 twelfth-graders strongly disagreed. Unless we view this difference in attitude as somehow due to rapidly changing times, we must conclude that, between the sixth and the ninth grade, in Oregon at least, something bad can happen to a child's attitude toward teaching.

Richey and Fox have something to add to this conclusion from their 1948 study of Indiana Freshmen. Among 69 5 students who had definite intentions never to become teachers, 20% had formed these intentions before high school, 93% before college. The most popular two-year period for this negative decision was in the ninth and tenth grades.

About additional implications of early decisions, the authors of the Yale-Fairfield Study say of their teachers- to-be:

> Those who began early to consider teaching less frequently doubted the wisdom of their choice, and they more frequently reported a social service motivation. They also reported more frequently that they had been influenced to choose teaching by favorable family attitudes and by their teachers.
>
> Those who began late to consider teaching more frequently reported doubting the wisdom of their choice; they more frequently reported being motivated by the beginning salary and the working conditions. Also more frequently, they reported being influenced by friends, and more thought of teaching as a temporary form of employment.

Why Teaching?

We have seen that the choice of teaching is an early one. We may now ask: "Why is teaching chosen?"

There are material incentives—scholarships and forgivable loans, salary, and improved working conditions. There are spiritual incentives—social prestige, moral duty, and personal satisfaction.

Material incentives appeal to the practical. They make it realistic to pursue teaching as a career. Spiritual incentives make teaching attractive emotionally. They make the selection of teaching as a profession not only sensible but spiritually fulfilling.

What is the relative strength of these incentives? When Valentine asked his British teachers in training what influenced them most at the time of entering the profession, they listed "liking for teaching" and "ideals" first more than three times as often as "economic desirability," "long vacations," or "Board of Education grants."

Thomas Ringness reported in the September, 1952, *Journal of Experimental Education* that among a group of one hundred Wisconsin Seniors in teacher training, "favorite interest" and "serve society" were rated higher as reasons for choosing teaching than "security," "attractive surroundings," "short hours," or "prestige."

In the December, 1948, *Phi Delta Kappan,* Ellis Hartford tells the same story about 207 Kentucky undergraduates. Fielstra's group of California undergraduates in education rated "to work with children," "to help youngsters learn and develop sound values," and "to grow myself" as more important than "desirable working conditions," "prestige," or "security."

And among the Yale-Fairfield group, both Seniors in teacher training and experienced teachers ranked "work with young people," "nature of the tasks involved in teaching," and "social service" ahead of "prestige," "working conditions," and "security."

How do incentives fare among high school students? In the late 1920s Frances Austin studied the reasons British adolescents gave for wanting to be teachers. From a group of 1,105 secondary-school children, she selected the 284 who gave teaching as their vocational choice. In the February, 1931, *British Journal of Educational Psychology* she reported that these children listed "fondness for school subject" first nearly two to one and rated "fondness for teaching and children" as of about equal importance to "good salary" and "easy."

Richey and Fox's total group of 3,905 Indiana high-school children named "opportunity to be of service" and "the chance to work with young people" as advantages of public school teaching twice as often as "personal prestige" and "provides a permanent job."

Thus high-school students in general as well as college students headed toward a career of teaching, reported that spiritual incentives outweigh material ones. The leading incentives were personal satisfaction and moral duty.

Do Teachers Have Influence?

What part do teachers play in the choice of teaching as a career?

Richey and Fox asked their high school students: "Which [person] has been of the greatest help to you in deciding the kind of work you want to do when you finish high-school?" Half named their parents, about 10% named a teacher. The question "greatest help in deciding," however, is biased against naming teachers as a source of vocational inspiration. In addition, among these high-school students only 6% actually wanted to become teachers. Within this 6% the proportion who named a teacher must have been a good deal higher than the over-all 10% figure. In Austin's group of British children, all of whom did want to become teachers, *25%* mentioned the influence of a teacher.

Among college -students studying education, the balance between the influence of parent and teacher shifts. In the Yale-Fairfield Study, Senior women were found to be influenced by their parents only slightly more than by their teachers. Senior men were found to be influenced by their teachers more than twice as often as by their parents. Willcox and Beigel found that among their first-semester Freshmen the example of a teacher was mentioned slightly more frequently than "family influence."

One of their students said:

> When I was in school I became very friendly with one of my teachers. She seemed to be the nicest person I had ever met and had all the attributes which I hope that I will someday have. That is why I chose teaching.

Fielstra's Sophomores and Juniors rated teachers twice as high as parents on a scale of importance in their decision to teach. Ringness' Seniors mentioned teachers 59% of the time and parents only 27% as influences in the choice of teaching as a career. One of his students, who went on to teach history, said of his own junior high school history teacher:

> From this man I can see the direct beginnings of my liking for American history, the course he taught and the course I am presently majoring in …. His method of teaching was not one of bored disdain, but one of virile interest, a form of interest which he had the unmistakable ability to transfer to his students…. I can't think of a teacher who had more lasting effect on me.

Ringness concludes that "the teacher as a recruiting agent for future teachers has been too much disregarded."

Finally, a group of forty-one women preparing for elementary-school teaching here at the University of Chicago were asked who had most influenced them to become teachers. Sixty-four per cent named a teacher, while only 15% named a family member.

We can conclude that at least a quarter and perhaps two-thirds of the college students preparing to become ~ teachers ascribe their choice primarily to a teacher. We may be convinced that, among such students, the influence of teachers on the choice of a career is substantial. But we do not know as yet whether this influence is any greater than, or any different from, that on other students who are not studying education and who do not plan to become teachers.

May Seagoe compared a group of 122 California Juniors and Seniors who were not studying to be teachers with a group of 122 who were. She succeeded in matching the two groups so that there were only negligible differences between them in the number of men and women, college year, marks, and background. Each student was asked, "Have you ever wished to be just like a teacher?"

In the May, 1942, *Journal of Educational Research* she reports that the teachers-to-be marked half again as many teachers "they wished to be just like" as those who did not plan to become teachers. The teachers-to-be also marked a teacher "in elementary school" more than twice as often and the answer "never" half as often. Seagoe concludes that the importance of the classroom teacher of today in determining the teaching personnel of tomorrow cannot be over-emphasized.

Teachers-to-be are influenced more by their own teachers than are other college students. Are prospective teachers also influenced in a different way? Richey and Fox asked their Indiana Freshmen to consider all the people they knew best in the community and to compare public school teachers as a group with these individuals on twenty-four characteristics. The researchers compared the evaluations made by the one hundred Freshmen who definitely planned to become teachers with the evaluations by the 695 Freshmen who definitely were not planning on teaching careers. In general, the teachers-to-be evaluated teachers overwhelmingly more favorably than did the students who had definitely decided against teaching.

The discrepancies in the evaluations are particularly suggestive. There was virtually no difference between the two groups on the items "practicality" and "refinement." But on the items "magnetism," "leadership," and "good sportsmanship" the teachers-to-be evaluated teachers twice as favorably as those who planned not to teach. In addition, while the never-to-be group put "industriousness" first and "culture" fourth, the teacher-to-be group put "culture" first and "industriousness" eighth.

Thus the two groups held quite different images of teachers. The image among those who planned to teach was that of a strong but human teacher -magnetic, cultured, a good sport, a leader. The image among those who planned not to teach was that of a busy but mechanical teacher-practical, industrious.

Richey and Fox did not draw these implications from their data, but we dare not overlook them. The image a college student has of the teacher is the product of his years of schooling, of the way he was taught. If a human image is associated with choosing while a mechanical image is associated with avoiding a teaching career, then we must be very careful how we handle mechanical solutions to the classroom problem.

Some Influential Teachers

The group of women studying to be elementary-school teachers at the University of Chicago also described the person who made the difference in their choice of career. Their answers enrich our insight into the complex and far-reaching influence a teacher can have.

The interest in teaching can start early, even before a child realizes it.

> The person who influenced me most towards teaching was my fourth-grade teacher. I was not conscious of this at the time. However, when I look back I realize it was my fourth-grade experience which steered me towards teaching. The atmosphere in the room was very friendly and pleasant, and I really enjoyed going to school. This was due to the teacher's personality. She was a very friendly and outgoing person. There were other teachers along the way who intensified my desire to go into teaching. But it was my fourth-grade teacher who was the first one to do this.

Nor has the teacher necessarily failed if his student leaves high school without a conscious intention to become a teacher.

> My high school chemistry teacher was one of only a few of my teachers who were really interested in the children they were working with. A number of us became close friends with him and his family. He was also unusual in seeming to enjoy his work. Although I had no interest in teaching during high school and even had negative feelings on the subject, I feel that this teacher helped make it possible for me later to consider teaching as a profession.

Even when a student fails to become enthusiastic about the subject matter of a course, the force of identification with his teacher may shape his professional career.

> The person who most influenced my decision to teach was a high school science teacher. Freshman science was a compulsory course and since my interests lay in the humanities, I dreaded taking science. But somehow this teacher managed to arouse an enthusiastic response. to science and to himself. His own enthusiasm generated mine.
>
> If I had to evaluate his science teaching, using myself as a result, I'd regretfully think he had failed. Although I did a prodigious amount of work in science that semester, I never again resumed this interest. But perhaps he did not actually fail since my growth of self-confidence and interest in teaching may be even more worthwhile than an enduring interest in science itself.

The evidence on early decisions should not be construed as a sign that the end of high school is always too late to start a young person on the road to teaching. The choice of teaching can first become crystallized at the end of high school.

> The person who most influenced me to choose teaching was an English teacher I had in my last year of high school. I had enjoyed and admired other teachers before, but it was she who most reinforced my adult conception of the type of profession teaching was and aroused my interest in teaching as a profession I might enjoy. She was the type of person who was able to convey what she knew without being pompous but rather in a way that created further enthusiasm in you about the topic of which she was speaking. Though the classroom atmosphere was not loose in any way, there was a relaxation that came from being really interested in the topic.

How can an interest in teaching be aroused? One way is by inspiration.

> My seventh-grade math teacher seemed to me to be all that a teacher should be. She was strict but she had a sense of humor and she loved her subject. She made me love math for the first time in my life. A teacher can cause a student to enjoy almost any subject if she enjoys it and teaches it with enthusiasm.

Another way is to set a good example and then give the student a genuine chance to emulate it.

> The high school librarian directed me toward the teaching profession. She was the person I worked for while in high school. I was an assistant in the library. I had a great many opportunities to watch her at her job and it occurred to me that perhaps teaching would be an interesting profession for me too.

Advice, patience, and assistance can make the difference.

> My high school civics teacher directed me toward the teaching profession. He was very willing to take time and talk to seniors who were interested in further education. He showed me the opportunities there were for winning scholarships and with his aid I applied and received one. The friendship that was established between us during my senior year is still maintained.

Not to mention sincere interest and honest propaganda.

> The head librarian and English teacher introduced me to the wonderful field of reading. They told me the books I should be reading and they were interested in my reactions. They also influenced me to enter the teaching profession by giving me all sorts of information on this subject and by continuously telling me of the satisfaction and benefits of this field.

Twenty-six of these forty-one women ascribed their choice of teaching to teachers. Eighteen of these were attracted by a teacher's masterfulness- knowledge of subject matter, high expectations, and strictness. Only eight were attracted by a teacher's permissiveness- warmth, sympathy, and indulgence. Couple this emphasis on masterfulness with the importance of the human image of the teacher and you have the teacher who does most to create young teachers among his students-the masterful but human teacher. Students testify to the importance of this kind of teacher.

They want their teachers to help bring out the best in them.

> My high school English teacher had a philosophy of really expecting a lot from his pupils. No A's were given during the marking periods of the semester so that an A as a final grade was quite an honor.

> He was not the usual insipid type of personality out of class or in. I liked him for that and still do. This man was the first teacher who ever made me exert myself. He taught me that education is a do-it-yourself process.

Students want their' grades to stand for something, even if that means not always getting an A.

> The strongest impression of a teacher in my mind is of my high-school chemistry teacher. My particular experience with him that is outstanding is when he gave me a B in chemistry. I had been a straight A student all through high school, and I had the feeling that some teachers were giving me A just in order not to disturb the pattern.

Most of all, students want to be treated with respect, even if it means foregoing the cozy but smothering swaddling clothes of tender loving-kindness.

> My civics teacher was a complete dictator, strict and dogmatic. He was very intelligent and demanded top work from me. He practically forced me to study because I was afraid not to. However, I admired this man more than all my other high school teachers. He treated me as an intelligent student.

Barbara Czanko gives a rousing description in her article "A Teacher and Then Some," in the March, 1959, *Clearing House.*

> Miss Berg, a tall but graceful woman, possessed that rare quality which enabled her to maintain close feeling with her students without losing her control over them. It was not often that she had to raise her voice to a class, but even when she did, it was not a piercing scream or shout, but a voice firm with confidence and only slightly louder than her normal speaking voice. Too often when a teacher raises her voice, a mere straightening of spines in chairs is seen. However, when Miss Berg did so, it was not spine straightening that was seen; instead, it was a straightening of minds that was *felt!* She seemed to realize that she must get our minds back to work, not our backs straight in our chairs. Surprising as it may seem we respected her for this ….
>
> This woman I can never forget, for the impression she made on me is a lasting one. As I prepare to enter her field, I hold high this thought: "I will never be satisfied with myself as a teacher until I feel I have become as good as Miss Borghild Berg."

The Classroom Problem

The choice of teaching comes early. Spiritual incentives are dominant. Teachers have influence. What does this mean for the teacher shortage?

Solutions to the classroom problem fall into two groups: those that emphasize mechanical aids like television and teaching machines and those that emphasize human aids like teacher assistants and clerks.

Mechanical solutions focus on the transmission of knowledge. They propose the wholesale use of machinery to take the place of missing teachers. The object is to maintain the quality and intensity of transmission in spite of a decreasing proportion of human teachers.

Human solutions ·emphasize the division of labor. They propose the recruiting of manpower from untapped human resources. The object is not to get by with fewer teachers but to reorganize so that a wider variety of people can help with the work.

If we are wise, we will use the best of both proposals, but the emphasis, the guiding spirit, will be vitally important. We must consider the effect each solution may have on the critical career problem.

The original teaching machine went into mass production about five hundred years ago. The invention of the printing press and the manufacture of printed books opened new worlds to education. Through books, great words become available to everyone, everywhere. But have books ever replaced teachers? Have books solved a teacher shortage even once during the last five centuries? How can we expect more from our new machines?

By making universal education possible, by carrying to every man the seeds of intellectual curiosity, the excitement of knowledge, books aroused a desire for education. In this way books intensified the teacher shortage. Why should we expect less from our new machines?

The very machines now guaranteed to make our supply of teachers go further are going to multiply our need for teachers in the future.

Mechanical solutions to the classroom problem, like books, have everything to offer for the transmission and the preservation of knowledge. They also produce a danger. Mechanical solutions encourage the solitary consumption of knowledge, private absorption, self-centered application. But will the pursuit of solitude create effective teachers?

We know the inarticulate shyness of the person who has buried his life in books. Quintilian said:

> Let the future orator … become accustomed from his earliest years to face men unabashed and not grow pale by living in solitude and so to say in the cloister's shade. The mind requires constant stimulus and excitement, but in such retirement it either flags and rusts as it were in the gloom or else becomes swollen with empty self-conceit. For one who does not match himself with others must needs overrate his own powers.
>
> Then when he must display the fruits of his study, he gropes about in broad daylight and finds everything new and strange, as is natural with one who has learnt in solitude what has to be done amidst a throng.

The new machines, like books, will play a powerful role in lonely learning, in creating solitary scholars. The question is how can the machines produce effective doers? How can they produce social maturity? How can they create competent teachers?

This danger has an ironic implication. In the job of showing how to learn, how to live, how to teach—in the job of providing the human model with whom to identify—no machine can take the place of the human teacher. Unless we believe that mechanical teaching, programming a lesson, writing a textbook, can be done by specialists who have never learned their teaching person to person, teaching machines may cause their own ruin.

If future generations lean ever more heavily on mechanical teaching devices—if solitary learning becomes the rule—human teachers may become extinct. If they should, who will write new programs for the machines? Will the machines write

their own programs, or will we at last achieve divine perfection-our old programs flawless and no need to progress to new ones?

What effect do mechanical devices have on the incentive to become a teacher? The evidence shows that this incentive has an interpersonal origin early in life, that it flows from human experience with a human teacher, that it is spurred on by human personal satisfactions and human moral concerns. If mechanical devices intervene between teacher and child, between child and child—if mechanical devices decrease or undermine human contacts in school—they may kill off the next generation of teachers before they are born.

Early Teaching Experience

Suppose we use student aids. Suppose we do teach children how to help us teach. Will that increase the supply of teachers in the future? Will that alleviate the career problem? Let us examine the evidence on the effect of early teaching experience.

Don Orton asked 405 undergraduates in education what experiences caused them to want to become teachers. In the April, 1949, *Phi Delta Kappan* he reported that 58% named experiences in teaching. Orton concluded that "a great many college students are attracted to teaching because they have already had some first-hand experience with it."

Among Ringness' Wisconsin Seniors in teacher training 63% said they had had early teaching experiences and 58% had led school classes. One Senior said:

> Perhaps the greatest single factor in my choosing the physical education field as my major at the University was a direct outgrowth of my days in junior high school. I was on the school basketball team and the coach was having more work than he could possibly handle by himself. So he requested two gym assistants for each of his classes. I was accepted … I relished the responsibility … shared with the coach the burden of teaching the class. I was proud of this attainment.

We asked the forty-one women in teacher training here at the University of Chicago whether they had ever done any teaching before college. Eighty-one per cent of them gave us examples, and 56% of the examples were of an academic type of teaching.

As before we want to bring the evidence into sharper focus. What is the power of early teaching experiences to discriminate between those who do and those who do not want to become teachers?

Seagoe asked her students about their early teaching experiences. She compared the responses of the students who were preparing to teach with the responses of students who were taking other programs. When she counted the number of times a student said he had both taught and liked it the score in favor of the teachers-to-be in informal teaching experiences was: "played school," 80% compared with 60%; "cared for children outside your immediate family," 61% compared with 47%, and "camp counselor," 29% compared with 13%. When she turned to more formal teaching experiences, the score among the teachers- to-be compared with the others was: "taking charge of the class when teacher was absent," 65% as against 45%; "tutored a student in a subject," 51% as against 44%, and "taught a regular class," 15% as against 5%.

Richey and Fox found that 7 3% of their Freshmen who were for teaching had early teaching experiences as compared with only 46% of those who were against teaching.

Richey and Fox concluded that there was "a substantial relationship between the amount of experience of a teaching nature the students had had and the degree to which they were inclined to select teaching …. No other item of data in this analysis showed a clearer relationship with the tendency to want to become a teacher …. Administrators and teachers should make every effort to provide experience of a teaching nature for students in the public schools."

Some Children Who Taught

Early teaching experiences do indeed play a vital role in the decision to become a teacher. The women studying teaching at the University of Chicago described some of the early teaching experiences that influenced them. Some of the teaching was informal.

> When I was ten years old I organized a play club composed of about ten kids between the ages of three and seven. I used to take these kids to the park in the morning during the summer when their mothers wanted to have free time to shop or wash. With my younger brother's help we would teach these kids simple games, have them make different things out of construction paper, and things like that. The highlight was a puppet show we gave with puppets made from stuffed socks which had been discarded.

The play club is one natural beginning of a career in teaching Another is the role of informal tutor and counselor.

> I have a young cousin who has a handicap and comes to me quite often as a friend for help in schoolwork. Several times I have worked with him in arithmetic and spelling. He has bad some difficult social problems with the children of his own age in school getting adjusted to his handicap. I have talked to him for hours just as a friend so that he would have someone nearer to his age to come to for problems.

Experiences in a discussion club can play their part. One woman gave as a reason for her teaching:

> In high school I belonged to The International Relations Club and there students would alternately take over explaining certain political situations to the group. At this time, when I was in charge, the class was in my hands completely.

In school there are many opportunities for a student to try his hand as a teacher. One way is to help his own teacher by working with individual children. One student said:

> In grammar school I had a regular job taking care of one particular classroom. The class was second grade and I would later help with arithmetic and reading difficulties on an individual basis.

Another student said:

> Several times in elementary school and in junior high school my teachers chose me to help teach children who were very slow in their schoolwork. The most outstanding time was in

> ninth-grade algebra when I really built a good relationship with a slow student and helped him a little to understand the subject.

Another way for a student to try his hand as a teacher is to hold a small class of his own-with his teacher's support and encouragement but without his teacher's immediate presence.

> When I was in my first year of high school I taught a group of boys algebra during the lunch period. This was done at my math teacher's suggestion.

The student may assist his teacher with his regular class, taking the teacher's place for short periods.

> My science teacher gave me a great deal of opportunity to do demonstrations before the class and thereby keep a constant interest in the subject even though I was doing A work.

Experiences like these can begin early. They can form the basis for an incentive to teach and reinforce this incentive throughout the student's school career.

One of the Seniors in the Yale-Fairfield Study wrote on the back of her questionnaire:

> I received my inspiration to be a teacher in the third grade. My teacher frequently allowed me to lead the reading and spelling classes, a task I enjoyed very much. My mind was made up to be a teacher. As I got into the higher grades I liked school very much and definitely wanted to go to college. Throughout grammar school and junior high school I was given many more opportunities to lead the class. I found that I enjoyed getting up in front of them. Talking to teachers and friends, I was encouraged to follow my goal … I have never doubted my choice and have never thought of considering another vocation.

At the close of his Andover lecture in 1846 William Fowle said:

> The want of competent teachers is felt and acknowledged throughout our land, and great efforts are making to furnish an adequate supply. Although I believe teaching to be a natural gift, as much as poetry or music, still, like them, it is an art that must be studied and cultivated, and one that, perhaps, will be hidden, unless an opportunity is afforded for its exercise. Acquiring knowledge is not acquiring the art of teaching, any more than accumulating money is the same as active beneficence. Not one learned man in a thousand is able to communicate what he knows, clearly and simply, to a child. Practice is necessary; but few have this, until they are called on to instruct. How different is the case where children, as fast as they learn, are required to impart what they have learned to others. The truth is, that a well conducted Monitorial School is the best normal school in the world; for practice goes with precept every step of the way. If our common schools were conducted, even in part, on the monitorial plan those children who have any tact, any peculiar love or aptness for teaching, would soon show it; and who does not see that pupils thus brought out would furnish the very best stock for normal schools, and the demand for teachers would not only be supplied, but would be supplied with teachers of the true birth, born and bred to their business?

Is any solution to the teacher shortage more noble, more natural, or more practical?

Appendix C: On Behalf of a Personal Approach to Learning

Benjamin Wright

Abstract Reprinted with permission from the publisher, the University of Chicago Press, from: Wright, B. (1958). Educational news and editorial comment: On behalf of a personal approach to learning. *The Elementary School Journal, 58*(7), 365–376.

> It would seem that a book called *The Psychology of Learning* … should be of immediate relevance to the classroom teacher… Yet here is a good book on learning which, for all practical purposes, is likely to be of no more value to the teacher in the classroom than, say, a book on astrophysics would be to the mariner on the open sea.

This opinion of *The Psychology of Learning* by B. R. Bugelski[1] appeared in the *Elementary School Journal* for December, 1957. The reviewer, Jacob W. Getzels, goes on to question the relevance to education of many research findings reported from learning laboratories.

This is a problem that is distressing to me, too. When I study the latest texts on educational psychology and learning theory, searching for something to offer my classes, it is disconcerting to discover how much scientific research can be consummated without adding anything concrete to my knowledge of how children learn. When my students ask me, or when I ask myself, what we mean by "learning," I find myself in a quandary. I am faced, on the one hand, by authoritative learning theories, the practical application of which I cannot understand, and, on the other, by the difficulties of improvising an answer that draws on whatever of my own learning experiences I can understand.

As a teacher and parent, I need some idea of what the phenomenon called "learning" really is. I need something practical to say about how children learn. Indeed, I need something practical to say about how I myself learn. If an abyss yawns between the laboratory science of learning and the classroom art of teaching, I must find a way across the gap or I must carve out a different road of my own, a road that will take me toward a practical understanding of learning and teaching. The abundance of unsuccessful efforts to bridge this gap has convinced me that neither I, as a parent and teacher, nor the learning theorists are going to come together in the near future. Therefore, if I want to understand learning, I will have to find my own starting point and make my own way. It may help me to begin if I try to understand why the problem has been such a difficult one in the first place.

[1] B. R. Bugelski, *The Psychology of Learning.* New York: Henry Holt & Co., 1956.

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2

The Rise of Objectivity

The last two or three centuries have been distinguished by the flowering of physical science. The burgeoning has been the result of the diligent application of objectivity. In early times man tried to understand his world subjectively, that is, in terms of himself. When I watch my children or recall my own childhood this kind of first approach to understanding seems natural enough. ~Ian's early efforts in this direction, like a child's, were haphazard and impulsive. The history of primitive magic and religion abounds in interesting examples of this kind of sally toward understanding. But, objectively speaking, these efforts were somewhat ineffective. By contrast, the sensational success of objectivity in dealing with the non-human world left man disillusioned with the subjective approach to his quandaries.

What brought about this disillusion? What happened to undermine man's confidence in his subjective sense of his world? More important, what happened to undermine man's confidence in his subjective sense of himself?

A short history of the crumbling of man's trust in the subjective appears in "One of the Difficulties of Psychoanalysis," written in 1917 by Sigmund Freud.[2] Musing over the widespread resistance to his revolutionary views, Freud noted that his studies of the unconscious were one more in a series of painful blows to man's faith in his ability to explain his world effectively in terms of his subjective experience.

First, in the sixteenth century, Copernicus had upset man's conviction that his earth stood at the center of the solar system. The inexorable procession of sun and planets were better described by viewing the sun as the center of things. Freud called this news the "cosmological" blow to man's narcissism.

Then, in the nineteenth century, Darwin and his associates upset man's conviction that there had always been a gulf between him and all other forms of life, man's conviction that he had always been separate and unique. A more plausible and useful explanation of the varieties of life on man's earth and of man's own physiological structure, the Darwinians said, was an evolutionary process that extended even to man himself. This news Freud called the "biological" blow to man's narcissism.

Freud's contribution, of course; was his theory that man is not even master in his own soul. Freud showed the world that man's thoughts and behavior are subject to unconscious instinctual forces of which he is often uninformed and over which he seldom has full control. Freud's main purpose in writing "One of the Difficulties of Psychoanalysis" was to explain the resistance to his theories. They roused opposition, he wrote, because they dealt a "psychological" blow to man's narcissism. But in this paper Freud was dealing with more than resistance to his ideas. Actually, this pioneer in whatever science of subjective experience exists today, was chronicling the progressive discrediting of subjective experience as a useful source of information.

The rise of objectivity seems to have been appropriate for the sciences of man's world. But objectivity has not been so uniformly productive in the science of man himself. Although the biological sciences have done well with this approach, objectiv-

[2] Sigmund Freud, *Collected Papers,* Vol. IV. London: Hogarth Press, 1949.

ity may be unable to clear up the confusion in man's social sciences. I wonder whether our difficulties are not unnecessarily compounded by the pains we take to overlook that important source of information, man's own subjective sense of himself.

The Objective Study of Learning

Most efforts to create a science of learning have tried hard to follow faithfully in the objective footsteps of the physical sciences. It has been difficult to conduct objective learning experiments, using rigorous methods, on man himself. Still, several ingenious attempts have been made.

One has been the objective study of humans, usually college students, performing highly simplified acts, such as conditioned eye-blinking or the learning and unlearning of nonsense syllables and numbers. Perhaps we are fortunate that college students are willing to spend so much time in this kind of activity. The trouble is that the learning studied is not of a kind commonly encountered in everyday life. It is hard to understand what role these fragments of behavior play in a complete act of learning. The relation between nonsense syllables and education, for example, is not entirely clear.

What is the rationale for this approach? We are told that the experimental task must be as simplified and as uncomplicated by "extraneous" factors as possible so that the essential nature of the learning process may be revealed. Unfortunately this approach has yielded little up to the present. Maybe the "extraneous" factors so carefully excluded, far from being extraneous, are of the essence.

Another approach has been the study of animal learning. Many modern scientific theories of learning stem from studies of animals. What is the rationale for these studies? It is asserted that important essentials of man's learning are also found in the learning of other animals. This line of attack will prove useful only when it reveals features in the learning behavior of different species, including man, that are vital in the unique learning of each as well as common to the learning of all. Researchers studying animals have arrived at a few foregone conclusions and created that inscrutable abstraction, the learning curve. This seems to be as far as they have been able to push their search for common elements. When one reads their reports, it still seems that amoebas do not learn like fish, fish do not learn like chickens, chickens do not learn like rats, rats do not learn like monkeys, and monkeys do not learn like college students. The results of animal experimentation have produced a variety of intriguing scientific theories of animal behavior, but I cannot figure out how to use these theories to help me clarify, in any practical way, the puzzle of human learning.

When learning theories based on the study of animals are applied in the classroom, efforts to understand humans as if they were animals turn out to be based on efforts to understand animals as if they were humans. Just where this roundabout logic promises the most, it rests most heavily on the experimenter's subjective interpretation of human elements in animal behavior. This line of reasoning seems plausible enough, but why begin with animals? Why not begin with the experimenter's interpretation of human elements in human behavior?

The socialization approach represents another effort to study learning objectively. The study of human learning in terms of development, socialization, and the social context in which learning takes place has become widespread. In this approach, studies focus on children and their social life. A large number of hopefully precise and hopefully relevant measurements are made. The school and family life of these children is observed, and researchers try to relate the many measured variables one to another and then to an evaluation of the children's school and family life.

This approach is popular, and for good reason. It seems to be more promising than animal psychology. Here, at least, children, not animals, are being studied. But so far, the harvest of solid knowledge useful to teacher and parent from this quarter, too, has been disappointingly meager.

Possibly the socialization approach to learning, for all its promise, is premature. The approach suffers sorely from difficulties of definition. What is a measurement? What is a variable? What is a social experience? And, perhaps most puzzling of all, what is a child? The approach leans heavily on tests and measurements, but the scores obtained are not the same as the child to be understood. Researchers have trouble steering their way between the Scylla of irrelevance and the Charybdis of imprecision, the most treacherous temptation being the sacrifice of relevance for the sake of precision. The focus of study, even the interpretation of results, tend to become defined by the tests available. Most studies of intelligence, for example, have been limited to the analysis of test scores. Researchers find themselves acting as though intelligence were no more than a score on an intelligence test. Perhaps this is one reason why we know little about intelligence. In reading research, speed has been easy to measure and comprehension difficult to assess. This state of affairs has led to an experimentally created illusion. The illusion has come to be interpreted to mean that speed produces comprehension. In trying to understand children, we find that even the best scores show us only fleeting facets of only part of a child's behavior. The art of assembling multiple scores into a useful reproduction of the original child still escapes us. Perhaps more serious, we are not entirely satisfied that even all the child's measurable behavior tells his whole story.

Some say that socialization is the only place to begin to understand learning. The starting point, these researchers insist, must be social life, but the socialization theories of learning now available set me adrift on a sea of unanchored and disconnected complexities. It is painfully hard for me to understand the social life of children. Small wonder, since I so dimly understand the child himself. I would be content if I could begin to understand what goes on between mother and child. It is said that Einstein, when asked why it was that physicists made so much more progress with physics than chemists did with chemistry, replied, "The trouble with chemists is that chemistry is too hard for them." My trouble with the socialization approach to learning is that it is too hard for me.

A Subjective Approach

What, then, is a teacher and parent in search of a useful understanding of learning to do? One of my troubles is that any theory of learning that does not explain my own experience of learning will never seem plausible to me, whatever conscientious service I may try to give that theory. Freud devised a useful way of understanding dreams, largely by the careful and subjective study of his own dreams. Perhaps a useful science of learning is waiting to be developed by those who will take the time to study the vicissitudes of their own efforts to learn. Having gained insight into their own learning process, these subjectively oriented researchers can use the results of their self-study as a base from which to study the learning of others. Objectivity may be the royal road to reliable knowledge about the external world. But when we are trying to understand ourselves and how we learn, scientific objectivity does not seem to be enough. Perhaps we need to embark on another road, one that is more subjective.

With respect to the study of himself, man is in a unique position. This is the only area of inquiry in which man is able to be the subject as well as the object of his study. But the possibilities of this position for gaining knowledge about learning have been generally neglected. It is surprising that this fact of life, while receiving considerable professional attention since the popularity of Freud's work, has had only slight influence in shaping the design or evaluation of research in learning.

Occasionally the unavoidable impact of subjectivity in research is explicitly recognized. But then the influence is usually acknowledged only as a source of error. Efforts are focused on trying to rid the experiment of its subjective aspects in order to approach the hopefully scientific goal of objectivity. But these efforts at objectivity sanforize right out of the research the very data that, it seems to me, are most likely to help me out of my dilemma. Instead of trying our best to get rid of the subjective aspects of our research, we might better try our best to harness our subjective experience in a way that would allow us to sort out and make the most of its contribution.

Three Ways to Learn About Man

There are, it seems to me, three ways man can learn about himself. Occasionally, one way may be more pertinent than another, but, in any inquiry, all three play a part. If one way is really major, the other two may be troublesome. The proper goal may indeed be to try to gain control over their presumably less important contributions in such a way as to exclude them. But for most problems in the study of man, it is helpful if all three ways can be evaluated together.

An Objective Way

The first way to learn about man is the objective one of physical science. In this way, man studies other men as though they were quite different from him—like rocks or trees. He tries to describe and measure, as free from any personal bias on his part as possible, the actions of other men. He tries to establish his observations in such a form that others can confirm them by following similar procedures. Sometimes this criterion is defined less strictly. The researcher is expected to report procedures and findings upon which competent witnesses can agree. It is often thought that only a non-subjective approach can hope to meet the requirements of these criteria.

The data collected in the objective study of man are descriptions of overt behavior, such as physical measurements and test scores. Every effort is made to free these observations from the influence of subjective bias. Often the results are said to be truly scientific only when these efforts at objectivity are thought to have been successful. But, while such results are a good beginning, they barely scratch the surface of what we want to know about the nature of man.

The annals of projective techniques offer one example of the difficulties that arise when the effort is made to hold exclusively to the objective way of learning about men. Several extensive projects were inaugurated to put the projective tests on a firm objective footing. But the objective methods devised for scoring projective protocols proved to be disappointing. They fit their subjects poorly and often yielded only information that can be easily obtained conversationally without recourse to any tests. Fortunately the host of practicing clinicians did not wait for these Procrustean objective methods but continued to gain incisive and hard-to-come-by insights into their clients by subjective means.

Two Subjective Ways

There are two other ways by which man can learn about himself. I think competent witnesses can find ways to agree upon the evidence produced by these approaches. But since they are not approaches that are usually called “objective” and since I want especially to draw attention to their subjective quality, I will call them “subjective” ways to learn about man.

The first has to do with the emotional impact the person being studied has on the one who is studying him. Suppose we are studying the behavior of a child. We are describing his movements and recording his test scores. However objective our approach, a great deal is going on between the child and us that neither a test score nor an objective description of movement will show. The child’s actions and emotions are having their effect on us. They are nudging our feelings this way or that. We are responding emotionally to what the child is doing with us. These responses in us tell us something more about the child than we can learn from the objectively describable aspects of his outward behavior. If we can evaluate these feelings, which are a response to the child, we shall have a fuller view of him.

Teachers and parents can add to their understanding of a child by observing the emotional impact that the child has on them. Does the child make them happy or sad? Does the child make them angry or content? The most crucial questions of all, perhaps, are linked with anxiety. Does the child make the parent or the teacher anxious? In what way? About what?

When child and adult are together, what does the adult feel like doing for the child? The child who feels deeply inadequate, for example, often has a talent for making the adults around him feel like doing many things for him. He makes them feel like hovering over him, perhaps even to the extent of treating him as if he were more inadequate than he actually is. What the teacher can learn from his own reactions is how inadequate the child feels. He may also learn how the child gets even for feeling helpless by making his teacher his slave.

The child who feels persecuted often provokes the adults around him into feeling like persecuting him. A teacher may discover himself finding excessive fault with this kind of child and feeling guilty about it. From this reaction the teacher can learn something about the child. He can learn that the child is angry. He can also learn that the child is searching for a plausible rationalization for this anger by provoking adults into treating him in such a way as to earn it. Finally, the teacher may sense the even deeper need of the child to enlist his teacher's concern. The child shows him this the best way he is able by trying to make his teacher feel guilty for treating him badly. What the teacher might do under circumstances like these is another problem. The point I am trying to make at the moment is that the feeling created by the child in his teacher can be relevant and useful information about the child.

The final way by which man can learn about himself is also subjective. This approach calls for a special kind of inner act, which is performed by a person so that he may examine the feel of this act in himself j and see what it tells him. The rationale for this approach is that we are much like one I another. In this act, the student of man I equates himself to other men. He says in effect, "Since we look alike, we must also j feel alike." To learn about other men h~ then asks himself, "If I were acting as that person is, how would I feel?" Or, "Under what inner circumstances would I do what I see that other person doing?"

Suppose that a teacher is trying to fathom why one of the capable children in his class repeatedly hands in his homework late. The teacher explores the circumstances under which he might do the same thing if he were in the child's place. On the basis of such a subjective exploration, the teacher may wonder whether the child is asking for stricter limits. The teacher may decide to try supplying limits by enforcing the timetable. On the other hand, the teacher may sense, instead, that the child is trying to be sure of his freedom of action. This explanation may seem plausible to the teacher in light of what he knows about the child. In this event, the teacher may decide that the child's welfare is better served by relaxing the timetable than by insisting on stringent enforcement.

Whatever the teacher does, if he tries to understand the child in this way, he is less apt to react impulsively in terms of his own annoyance or mechanically in terms of school regulations or permissively in terms of lack of interest. Instead, the teacher may be able to use his insight into the child's feelings to plan a course of action that

includes a feeling for what moves the child. These, it seems to me, are the kinds of actions that offer the most promise for us and for our children in the classroom.

For most situations it suffices to ask, "If I were that other person, how would I feel?" But sometimes the problem facing us may be even more difficult. Perhaps we are trying to understand a child who is in emotional trouble. Then we may have to do more than try to imagine how he feels. We may have to put ourselves through some of the motions we observe him going through in the hope that the experience will affect us in some way as it does him. If we succeed, the experience we then have may give us some sense of what his inner world is like. The venture may be far from easy, particularly if his behavior is upsetting to us. But this is one of the best ways I know, for example, to get some notion of the inner world of the severely disturbed child.

There is nothing new about this way of learning about man. It is used every day. Insofar as we recognize each other as alike, we assume, quite without thinking about it, that we must *feel* alike, too. What has been missing in the study of learning is the explicit use of this kind of subjectivity.

Some Difficulties

The use of subjective experiences as scientific evidence faces several difficulties. First of all, we have learned to suppress many of our subjective reactions because we have often had the unpleasant experience of seeming to be wrong in our assumptions about how others feel. Sometimes we really *are* wrong. At other times, it only seems that we are wrong because the emotions that we have sensed are emotions that it is customary to deny. We do not like to risk the chance of a mistake when we are trying to be scientific. The alleged uncertainty of subjective methods troubles us. As a result, when we are striving for accuracy we tend to eschew this apparently chancy method for one we hope is more reliable.

Another difficulty is that what we want from the child and what he wants from us are not usually the same thing. Yet our subjective experience of our wants and his can easily become confused. We *feel* both sets of wants in the same place in ourselves. We have to devise strategies by which we can separate them.

Perhaps the most troublesome difficulty of all is that we are as much subject to, as we are masters of, our own state of mind. This difficulty takes two main forms. We do not expect our view of another individual to coincide at all points with his view of himself. Yet sometimes we see feelings in others that are, at most, barely there. At other times, we overlook the obvious and cannot see at all what is right in front of us. Why?

There are always some things about ourselves that give us discomfort—some part of the truth about ourselves that is so unpalatable to us that we cannot abide it. We try to relieve our uneasiness by excluding this unwanted truth from our picture of ourselves. One way we do this is by seeing these uncomfortable things in others instead of in ourselves. Psychologists call this mechanism "projection," and we all project at times. This fact complicates matters when we try to sort out and use our own subjective reactions.

Sometimes, if we do not want to acknowledge a feeling in ourselves, we may find it more economical emotionally to keep ourselves blind to it in others as well. This is the other way we try to exclude unwanted truths from our picture of ourselves. Psychologists identify at least two degrees of this behavior, which are called "denial" and "repression." What we do is to censor or distort whatever we do not want to see. The effect is to render unavailable, or at least unclear, emotional reactions in us that may be quite important to any effort to understand others.

Without a doubt, both of these difficulties—responding to feelings that are hardly there or being blind to feelings that are decidedly present—are handicaps in using subjective experience. Still we can deal with these difficulties. One way of doing so is to pool our impressions with the impressions of others who have shared our experience. It is by talking about our impressions that we discover in the first place that not everyone always agrees with us about them. We have all had the experience, for example, of discovering that someone else's impressions of a mutual acquaintance differ from our own. Usually what we do then is to defend our impressions staunchly or give in to the view of the other person, feeling chagrined, perhaps, for having been "wrong."

But neither alternative is constructive. Impressions of persons are alike in some respects and unlike in others. The fascinating question is how and why. It is normal for two people to have differing but equally relevant impressions of a third person. It is the exploration of such differences that can free each from his own distortions and enrich the insights of all into what makes this other person tick.

At the University of Chicago Orthogenic School, described so well in Bruno Bettelheim's books, *Love Is Not Enough*[3] and *Truants from Life*[4], a group of counselors and teachers work together around the clock with severely disturbed boys and girls. The children are baffling. Many come to the school only after individual treatment has failed to help them. Yet these child-care workers find ways. One of their basic tools is the daily staff conference in which counselors and teachers share with each other their experiences with the children. By talking to one another about their reactions to the children and by sorting out implications of their reactions for their individual efforts with each child, these workers are able to succeed where others have failed.

Two teachers talking over their impressions of a child they both work with have an unparalleled opportunity to use all three ways of knowing about that child. Each can help free the other of his projections and distortions. I do not mean that they should try to agree about the child. Nothing could be further from my point. I mean that by comparing what they see and what they feel, they have an opportunity to learn a great deal more about the child and about their relations to him than either could alone. The goal, far from being firm agreement, should be an exploration of the information each teacher can contribute in a joint effort to gain understanding.

[3] Bruno Bettelheim, *Love Is Not Enough.* Glencoe, Illinois: Free Press, 1950.

[4] Bruno Bettelheim, *Truants from Life.* Glencoe, Illinois: Free Press, 1955.

A Personal Approach

My troubles in searching for an approach to learning began when I found myself at a loss as to how to use the learning theories already available to help me understand learning. This distressing situation motivated me to try to understand what was wrong, and I ended up exploring various ways I might hope to learn about man. The customary objective approach seemed to be inadequate. A more subjective approach seemed called for. If there is any value in a subjective approach to understanding learning, then the best place for me to begin, I realize, is with myself. Further, when I turn to use whatever I discover about learning from myself to the study of others, I will want to begin by asking them to tell me what their learning feels like, not by administering tests. My approach to learning will have to be a personal one.

To begin with myself then, when I review recent events in my life that might be called learning experiences, I see two main kinds. Usually my learning has to do with the world around me. I read a book, ask someone a question, solve a problem. My attention is directed toward coming to grips in some way or other with the world. I am changed by such encounters, and I learn. But there is another kind of experience that leads to learning, too. This experience is often called "thinking." It has to do with what is already inside of me. I do not read a book, but I may search my memory. I do not ask someone else a question, but I may ask one of myself. Rather than explore the outside world around me, I explore what is going on inside. Let me call this learning by thinking "meditation."

Two Topics of Meditation

My meditations, I discover, have two main topics. The first is my recollection of the outside world. I review what I know about the world to see what conclusions I can draw, what plans for action I can make. My attention is focused not on the world but on my accumulated experience of it. I learn by exploring this experience. When I am through meditating, my thoughts and ultimately my behavior are changed, although I may have had no intervening commerce with the outside. What startles me about this ordinary everyday experience is that I can see no way to infer its existence in others except in terms of my having first experienced it in myself. How can I observe this experience in someone else by objective means only? So far I have found no way. Yet this experience seems a most important part of my learning, and I am convinced that it is important in the learning of others, too.

In its simplest form, this kind of meditation seems to be just figuring out what to do next. But this inner activity seems also to take place in less explicit and more complex ways. Scholars, for example, report how they puzzled over a problem for days, weeks, sometimes even for years and seemed to be getting nowhere. Then, when the problem seemed least on their minds, a solution suddenly came to them. Occasionally it is reported that solutions have come to people while they were asleep and dreaming. I believe these accounts, not because I can observe them

happening in others, but because I have had the same experience myself. The details of a problem that I cannot solve are on my mind. I do not seem to be making any headway. Then, often at a time when I am working on something quite different, to my surprise I suddenly see how the details fit together. The problem is solved. This kind of inner experience, which ends up as learning and which is not only unseen from the outside but also only indirectly or partially felt on the inside, can be puzzling. It seems to be an important part *of* learning. I do not see how we can find out about it except through a personal subjective approach.

The second topic of meditation is myself. True, this topic also reflects the impact of a long history of encounters with the world, but what I am trying to figure out now is not what my view of the world is like but what I am like. What I want to know is not what to do with the world but what to do with myself. We seem to be most familiar with this kind of experience in psychotherapy. In therapy a person rearranges his inner life in order to live more successfully with himself and with others. He accomplishes this change by exploring his inner life with the help of a therapist. His inner life is reified in the privacy of the therapeutic relationship, and inner structure is somehow altered by means of "working through." Therapy is often spoken of as a learning experience. It seems like a learning that takes place because of a special kind of meditation about oneself.

Can this kind of learning take place outside a professional therapeutic relationship? Can it be done without the help of another person? Psychotherapists differ. When the inner situation is a desperate one, it seems to me that outside help may be indispensable.

On the other hand, I, for one, often have less dramatic experiences when a modest exploration on my own of how I feel teaches me something about myself that is useful in deciding what to do. This feels like learning to me. It is a kind of experience that we all have and regularly put to good use in our daily lives. I do not understand what part this kind of experience plays in academic learning, but its weight in my own everyday life tells me that it must play a substantial and complex part.

There are two ways by which I learn about the world. Sometimes I learn by doing, by participating physically, by being the one who takes the risks. When I do this, my learning tends to follow closely the form of the experience in which I am participating. It tends to be almost as specific, concrete, and limited in scope as that situation. I may generalize on this experience later, but that will be another event in my education, perhaps one of the meditative kind described earlier. This participant learning has to do with the here and now. The gain is tangible. The risk is high. It seems to be a good way to develop discipline and skill. The second way I learn about the world is more vicarious. It tends to be a low-risk, diffuse-gain way of learning which can be quite abstract, and broad in scope. It can proceed leisurely and seems made more for future reference than for immediate use. This vicarious way of learning seems to be important in expanding knowledge and wisdom. While it would be hard to maintain a sharp distinction between these two ways of learning, since both play their part in real experience, it is useful to me to think of them as different in emphasis.

If I go to boot camp to learn the life of a marine, if I sit down to practice the piano, my experiences lead to participant learning. But when I watch a baseball

game, read about Byrd's adventures at the South Pole, study the philosophy of Dewey, my experiences lead to vicarious learning. As a participant, I tend to confine myself to the situation at hand. The room I have for trial and error or for exploration is thus limited. As an observer I am missing the hard test and the immediate tangible rewards of reality, but I can have all the trial and error I have imagination for, and I have wide reaches of time and space in which to roam.

Some things can be learned best directly, others vicariously. Sometimes the requirements of participation may prevent the very exploration of alternatives that is a special province of academic education. The life-adjustment curriculum in the community-centered school, for example, may have some unrecognized drawbacks along this line. To focus the child's learning on adjustment to an immediate community, particularly if it is a homogeneous one, may have a limiting, as well as a stimulating, effect. To require the child to invest most of his energy in an exhaustive adjustment may leave him less energy to explore a variety of rewarding alternatives.

It helps me to make a distinction between relevance and risk here. Risk has to do with whether I am participating physically and thus learning by doing or whether I am looking on and thus learning vicariously. Relevance has to do with how close a meaning the subject of study has for me. When I am participating, the experience has both relevance and risk. When I am looking on, the risk is less, though the subject of study may still be highly relevant. When a child studies the way his teacher treats another child in order to learn what kind of a teacher he has, the child's risk may be low, although the situation is highly relevant to him. If the child were himself the one involved with the teacher, his experience might have only a little more relevance, though it would certainly have much more risk. The impact of this added risk often interferes seriously with the child's ability to think clearly. The high risk prevents him from being able to learn what kind of teacher he has.

But this is only a beginning. A personal approach to learning leads to other topics. One is the interplay between participant and vicarious learning in any actual learning experience. Another is the puzzling nature of creative thinking and the role of mistakes in learning. While we emphasize getting the right answer in our schools and on our tests, for example, we have all had the experience of learning more from our mistakes than from our right answers. We know that some of our most learned men were failures at getting right answers early in life. Finally, the main topic to which a personal approach to learning leads is interpersonal relations and their central role in learning, including the vicissitudes of that profound and mystifying phenomenon called "identification." This is the place where the complexities arise with which the socialization approach to learning tries to deal.

All I have done here is to limn an introduction to some learning phenomena, which I do not know how to study objectively. What little I know about them depends on personal experience with them in myself. I can infer their existence in others by listening to what they tell me and by thinking in terms of our basic similarity to one another. That is why I am wondering if a more personal and subjective approach will not be helpful in understanding what human learning is all about.

Appendix D: List of Dissertations as Supervisor and Committee Member 1958–2001

Benjamin D. Wright

Ph.D. Dissertations: Chair

Thomas O'Neill (Spring 2001) Explaining Rating Scale Usage: The Semantic Threshold for Induced Categories, University of Illinois at Chicago.

Brian Bontempo (2000). Assessing Speededness using Probabilistic Models. University of Chicago.

Louise White (Spring 1998) Equating Low Back Pain, University of Illinois at Chicago, Department of Physical Therapy.

George Karabatsos (August 1998). Analyzing Non-Additive Conjoint Structures, University of Chicago.

Stuart Luppescu (Autumn 1996). Virtual Equating: An Approach to Reading Test Equating by Concept Matching of Items, University of Chicago.

Winifred Anne Lopez (Summer 1996) The Resolution of Ambiguity an Example from Reading Instruction, University of Chicago.

Richard C. Gershon (Spring 1996). The Effect of Individual Differences Variables on the Assessment of Ability for Computerized Adaptive Testing, Northwestern University (secondary advisor).

Mark H. Moulton (Spring 1996) N-Dimensional Replacement Implications of a Rasch Geometry, University of Chicago.

Gregory Ethan Stone (Winter 1996) Criterion Referenced Standard Setting, University of Chicago.

Yi Du (Autumn 1995). Measuring Student Writing Abilities in a Large-scale Writing Assessment, University of Chicago.

Sunhee Chae (Spring 1995) Item Equivalence from Paper-and-Pencil Computer Adaptive Testing, University of Chicago.

Katarzyna C. Szydagis (Spring 1995) Quantifying Self Psychology for African-American Students, University of Chicago.

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2

Linjun Shen (Spring 1994). Assessing General Medical Knowledge, University of Chicago.

Anna K. Bersky (Winger 1994). The Validity of a New Test of Nursing competence, University of Chicago.

David Zurakowski (Spring 1993). The Structure and Growth of Human Intelligence, University of Chicago.

Patrick Fisher (Winter 1993; MA thesis) Measuring Baseball Performance with Rasch Analysis, University of Chicago.

Sandra Dolan (Winter 1993) A Comparison of Computer Adaptive Test Administration Methods, University of Chicago.

Ong Kim Lee (Autumn 1992). Measuring Mathematics and Reading Growth, University of Chicago.

Bahrul Hayat (Autumn 1992). A Mathematics Item Bank for Indonesia, University of Chicago.

Betty Bergstrom (Autumn 1992). Computer Adaptive Versus Pencil and Paper Tests, University of Chicago.

Eunlim Chi Kim (Summer 1992). Factors Affecting the Difficulty of Phoneme Identification: The Case of Korean Children Learning ESL, University of Chicago.

William John Boone (Autumn 1991). Improving Elementary School Science by Application of Item Calibration Mapping, University of Chicago.

Anthony James Pitruzzello (Summer 1991). Measuring Social Desirability Response Bias, University of Chicago.

Margaret McCabe, CAS, (Summer 1991). Evaluating the Validity and Reliability of the Pediatric Functional Independence Measure, Rush University.

Donna Surges Tatum (Spring 1991). A Measurement System for Speech Evaluation, University of Chicago.

Anne Louise Wendt (Winter 1991). Clinical Environments and Student Attitudes Toward Mental Illness, University of Chicago.

Barbara Jean Davis (Spring 1990). Perfectionistic Thinking in Teachers. University of Chicago.

Nikolaus Bezruczko (Spring 1990). The Construction and Validation of a Rasch Preference Scale for Design Simplicity: An Aspect of Aesthetic Judgment, University of Chicago.

Judith A. Beto (Spring 1990). Self-regard in Hypertension: A Study of Selected Quality of Life and General Attitude variables of Hypertensive Patients, University of Chicago.

Carol Monroe Myford (Autumn 1989). The Nature of Expertise in Aesthetic Judgment: Beyond Inter-judge Agreement, University of Chicago.

Wendy Lee Hick-Rheault (Autumn 1989). Learning Styles of Physical Therapy Students, University of Chicago.

Raymond John Adams (Autumn 1989). Estimating Measurement Error and Its Effect on Statistical Analysis, University of Chicago.

Wan Mohd Raru Bin Abdullan (Summer 1989). The Effects of Teacher Attitudes Toward Students on Teacher Planning, Instructional Support, and Teacher's Efforts in Maintaining Order in the Classroom, University of Chicago.

Lih-Meei Yang (Spring 1989). Medical Career Attitudes: Differences Among Specialties and Changes Over Time, University of Chicago.

John Michael Linacre (Spring 1989). Many-facted Rasch Measurement, University of Chicago.

Dorthea Juul (Spring 1989). Measuring Medical Problem Solving, University of Chicago.

Robert Charles Froh (Spring 1988). Improving the Information Quality of Student Ratings of College Instruction, University of Chicago.

William Paul Fisher, Jr. (Spring 1988). Truth, Method, and Measurement: The Hermeneutic of Instrumentation and the Rasch Model, University of Chicago.

Matthew Schulz (Autumn 1987). Functional Assessment in Rehabilitation: An Example with the Visually Impaired, University of Chicago.

Jennifer Frens Bosma (Winter 1985). Teacher and Student Responses to a System of Rational Measurement, University of Chicago.

Mark Wilson (Spring 1984). A Psychometric Model of Hierarchical Development, University of Chicago.

Larry Houston Ludlow (Autumn 1983). The Analysis of Rasch Model Residuals, University of Chicago.

Richard M. Smith (Autumn 1982). Detecting Measurement Disturbances with the Rasch Model, University of Chicago.

Anthony G. Kalinowski (Autumn 1982). Chronic Pain and Suffering, University of Chicago.

Michael Louis O'Brien (Summer 1982). Calibrating Item Difficulty as the Basis of Prescriptive Test Theory, University of Chicago.

Geofferey Norman Masters (Winter 1980). A Rasch Model for Rating Scales, University of Chicago.

Diana Krakower Calica (Winter 1980). A Study of the Relationship between the Cloze Test and a Hierarchical Model of Reading Comprehension Skills, University of Chicago.

Thomas Gene David (Winter 1979). The Assessment of Functional Properties of Classroom Physical Environments, University of Chicago.

Ronald J. Mead (Autumn 1976). Assessment of Fit of Data to the Rasch Model through Analysis of Residuals. University of Chicago.

Graham A. Douglas (Summer 1975). Test Design Strategies for the Rasch Psychometric Model, University of Chicago.

Charles James Nier (Spring 1975) Some Relationships between Psychological Structure: Educational Beliefs and Teaching Strategies in Three Types of Teacher Trainees, University of Chicago.

John Douglas Eggert (Spring 1975). A Multidimensional Approach to Assessment of affective Change in the Classroom, University of Chicago.

Charles E. Mosley (Autumn 1973). Race and Sex in Teacher-Pupil Relationship, University of Chicago.

Rosemary Likey Hake (Autumn 1973). Composition Theory in Identifying and Evaluating Essay Writing Ability, University of Chicago.

David Andrich (Autumn 1973). Latent Trait Psychometric Theory in the Measurement and Evaluation of Essay Writing Ability. University of Chicago.

Raymond Howard Comeau (Spring 1973) Some Relationships between Teacher and Student Personality Types, University of Chicago.

Julia Jane Hereford (Summer 1971) Self Concepts and Childhood Recollections of Undergraduate Women Preparing for Nursing or Teaching, University of Chicago.

Vanna Thorman Magsino (Spring 1971). An Inquiry into the Psychology Aspects of Truancy, University of Chicago.

Solomon Rockove (Winter 1971). Toward the Development of a Theory of Matrism and Patrism. University of Chicago.

Nargis Panchapakesan (Spring 1969). The Simple Logistic model and Mental Measurement. University of Chicago.

Bruce Choppin (Summer 1967). A Psychological Analysis of Linguistic Behavior, University of Chicago.

Sister Mary A. Stozek (Summer 1966). Self-concept Systems of Adolescents Planning to Become Teachers. University of Chicago.

Robert J. Panos (Winter 1966). Developmental Patterns in Student Teachers' Attitudes. University of Chicago.

Herbert Walberg (Winter 1964). Dynamics of Self conception During Teacher Training. University of Chicago.

Shirley A. Tuska (Jenks) (Summer 1963). Self-conception and Identification among Women Planning and Not Planning to Teach, University of Chicago.

Barbara Sherman (Spring 1962). A Study of Teachers' Identifications with Childhood Authority Figures. University of Chicago.

Ph.D. Dissertations: Committee

Helen P Makris (Winter 1999). Educational Resilience Mediating Factors of Adolescent Adversity, University of Chicago.

Jennifer Schmidt (Winter 1998). Exploring the Role of Action, Experience and Opportunity in Fostering Resilience among Adolescents, University of Chicago

Jaekyung Lee (Spring 1997) Multilevel Linkages between State Policies and Educational Outcomes: An Evaluation of Standards-based Education Reform in the United States, University of Chicago.

Rita Bode (Spring 1996). The Effect of Ability Grouping on Student Math Achievement, University of Illinois at Chicago, Educational Psychology.

Marta Elena Alvarado (Summer 1996). Psychosocial Variables which Affect Performance in Medical School, University of Chicago.

In-Soo Choe (Summer 1995). Motivation, Subjective Experience, Family and Academic Achievement in Korean High School Students, University of Chicago, Human Development.

Livia Magalhaes (Summer 1995) The Assessment of Motor and Process Skills during Naturalistic Classroom Observations, University of Illinois at Chicago, Occupational Therapy.

Samuel Whalen (Spring 1993). Challenge and Talent Development During Adolescence, University of Chicago.

Jian Zhang (Spring 1993). Statistical Significance Publication Bias: Its Determination and Statistical Adjustments in Meta-analysis. University of Chicago.

Kenneth Aaron Frank (Autumn 1993). Identifying Cohesive Subgroups. University of Chicago.

Yoshi Spencer DeRoos (Winter 1993). Short-Term Agency Based Training of Adult Day Care Staff. University of Chicago.

Albert Wallace Lyons (Summer 1988). Role Models: Criteria for Selection and Life-cycle Changes, University of Chicago.

George Engelhard, Jr. (Spring 1985). The Discovery of Educational Goals and Outcomes: A View of the Latent Curriculum of Schooling, University of Chicago.

Patrick Leo Mayers (Winter 1978). Flow in Adolescence and Its Relation to School Experience. University of Chicago.

Robert Edward Draba (Winter 1978). The Rasch Model and Legal Criteria of a "Reasonable" Classification. University of Chicago.

Harold Pates (Autumn 1976). Condescension: A Study of Attitudes of Teachers Who Work with Children in all Black Schools, University of Chicago.

Jasmin Espiritu Acuna (Autumn 1976). Opportunity Structure and Cognitive Growth, University of Chicago.

Randall Morris Johnson (Spring 1975). The Development of Instructional Activity for Urban Students Based upon Learner Defined Concerns, University of Chicago.

Lorraine Elise Granieri (Spring 1975) An Investigation of the Effects of Motives and Attitudes on Intention to Continue Foreign-language Study, University of Chicago.

Michael I. Waller (Summer 1973). Removing the Effects of Random Guessing from Latent Trait Ability Estimates, University of Chicago.

Stephan Harth Wilson (Winter 1972). A Participant Observation Study of the Attempt to Institute Student Participation in Decision Making in an Experimental High School, University of Chicago.

Trude Unger (Winter 1972). The Influence of Student Behavior and Teacher Personality on Teacher Behavior, University of Chicago.

William James Bramble (Summer 1971). Sequential Testing of Models for the Analysis of Covariance Structure, University of Chicago.

Gregory Arthur Hancock (Spring 1971). Public School, Parochial School: A Comparative Input Output Analysis of Governmental and Catholic Elementary Schooling in a Large City. University of Chicago.

Helen Hughes (Summer 1970). Variables Associated with Later Neuro-psychological Outcome in Children of Very Low Birthweight, University of Chicago.

John E. Hutchison (Summer 1969). The Subject Matter Specialist: Expectations Held Toward His Role. University of Chicago.

Clarence Bradford (Summer 1968). An Examination of Some Models and Techniques for the Analysis of Complex Systems in Educational Research. University of Chicago.

Emil Jost Haller (Autumn 1966). Teacher Socialization: Pupil Influences on Teacher's Speech. University of Chicago.

Br. Leonard Courtney (Spring 1964). The Relationship Between the Oral and Silent Reading of College Students, University of Chicago.

Eva Lenore Goble (Winter 1964). The Participation of the Young Homemaker in Group Learning Activities, University of Chicago.

Douglas E. Stone (Summer 1962). A Methodological Approach to the Analysis of Teacher Behavior that Reveals the Stability of Human Characteristics, University of Chicago.

Everett Arthur Johnson (Summer 1962). The Leader Behavior of Hospital Administrators, University of Chicago.

Marvin A. Brottman (Summer 1962). The Administrative Process as Perceived in the Behavior of the Elementary School Principals, University of Chicago.

Elizabeth Zimmerman Howard (Autumn 1961). Teacher Training and Student Change: An Analysis of Needs, Attitudes, and Performance, University of Chicago.

Donald Walter Peterson (Spring 1961). Prospective Teachers' Concepts of Self, Teacher, and School, University of Chicago.

Walter Johnston Hartrick (Spring 1961). Perceptions of Task and program of the Public High School, University of Chicago.

Irma Theobald Halfter (Spring 1961). The Comparative Academic Achievement of Women Forty Years of Age and Over and Women Eighteen to Twenty-five Years of Age, University of Chicago.

Gaber Abd El Hamid Gaber (Spring 1961). Needs and Values of Egyptian and American Secondary School Teachers: A Cross-cultural Study, University of Chicago.

Ramon Reyes López (Winter 1961). A Study of Attitudes Toward the Army among Male High School Seniors and the Relationship between these Attitudes, Social Class, and "Dominant Interests in Personality", University of Chicago.

George Henry Daigneault (Autumn 1960). The Arts Department Chairman as a Source of Role Conflict, University of Chicago.

Agnes Rezier (Autumn 1960). Needs, Perception, and Level of Aspiration in College, University of Chicago.

James Varnes Pierce (Autumn 1960). Non-intellectual Factors Related to Achievement in Above Average Ability High School Students, University of Chicago.

Martin Nichols Chamberlain (Autumn 1960). The Professional Adult Educator: An Examination of His Characteristics and the Programs of Graduate Study which Prepare Him for Work in the Field, University of Chicago.

Maurice Alan Brown (Autumn 1960). The Relationship of the Quality of Collegiate Education to the Continuing Education of College Alumni, University of Chicago.

John Morton Bahner (Autumn 1960). An Analysis of an Elementary School Faculty at Work: A Case Study, University of Chicago.

Joseph Soffen (Spring 1960). Training of Non-professional Leadership in Adult Education, University of Chicago.

Allen T. Slagle (Summer 1959). The Task of the Public School as Perceived by Occupation and Age Sub-publics, University of Chicago.

Roger C. Seager (Summer 1959). The Task of the Public School as Perceived by Proximity Sub-publics, University of Chicago.

Myles Friedman (Summer 1959). Conflicts in Learning, University of Chicago.

Roderick F. McPhee (Spring 1959). The Relationship between Individual Values, Educational Viewpoints, and Local School Approval, University of Chicago.

Lawrence William Downey (Spring 1959). The Task of the Public School as Perceived by Regional Sub-publics. University of Chicago.

Merton Verdell Campbell (Autumn 1958). Self-role Conflict among Teachers and Its Relationship to Satisfaction, Effectiveness, and Confidence in Leadership, University of Chicago.

Appendix E: Benjamin Drake Wright—VITA

Abstract This is the content of Ben's CV as it was found in late 2000, just after his health failed. From the entries in the CV, it appears Ben was in the midst of editing it and bringing it up to date. The unnumbered entries, and those out of sequence, are shown here where Ben left them.

Education and Certification

1939–1944	**The Hill School**, Pottstown, Pennsylvania Scientific Diploma, June 4, 1944 Cum Laude Society, May 22, 1944
1944–1947	**Cornell University**, Ithaca, New York Bachelor of Science with Distinction Physics, June 16, 1947
1947–1949	**University of Chicago, Department of Physics** Graduate work: Physics, Mathematics Sigma Xi Society, March 3, 1949
1948–1951	**University of Chicago, Committee on Human Development** Graduate work: Clinical Psychology, Personality Theory Doctor of Philosophy in Human Development June 7, 1957
1951–1954	**Chicago Institute for Psychoanalysis** Certificate in Psychoanalytic Child Care June 14, 1954

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2

1959	**The Board of Examiners** **Illinois Psychological Association** Certified Psychologist Certificate Number 155, April 9, 1959
1964	**Department of Registration, State of Illinois** Registered Psychologist Certificate Number 72–140, April 4, 1964

Employment

1944–1946	**United States Navy** Officer Training, USNTS Ithaca, USNH Sampson, Honorable Discharge, June 15, 1946
1947	**Bell Telephone Laboratories** Murray Hill, New Jersey Research Physicist Supervisor: **Charles H. Townes** (Nobel Laureate) (Microwave absorption spectra of iodine monochloride)
1947–1948	**Gads Hill Center**, Settlement House, Chicago Group Worker Supervisor: **Bernice S. Morrison** (Directed group theater for young adults)
1948–1950	**University of Chicago, Department of Physics** Research Physicist Supervisor: **Robert S. Mulliken** (Nobel Laureate) (Ultra-violet absorption spectra of organic molecules)
1950–1957	**University of Chicago, Orthogenic School** Counselor, 1950–1952 Supervisor: **Bruno Bettelheim** (Residential child care of emotionally disturbed boys) Psychotherapist, 1951–1957 Control Analysts: **George Perkins M.D., Anne Benjamin M.D.** (Psychotherapy with schizophrenic children) Research Associate, 1952–1957 Supervisor: Bruno Bettelheim USPHS Project M-476: Staff Problems Met in Children's Institutions (Research design, interviewing, test construction, factor and variance analysis, annual progress reports) Wieboldt Foundation Project: Treatment of Childhood Schizophrenia (Observation and treatment of schizophrenic children, life histories, semi-annual progress reports)

1957	**University of Chicago** **Departments of Education and Psychology** Instructor, 1957 Assistant Professor, 1958–1961 Associate Professor with tenure, 1962–1966 Professor, Education and Psychology, 1967–present Director, Education Statistics Laboratory, 1958–1966 Editor, School Review, 1969–1977 Chairman, MESA Special Field, 1979–1987 Director MESA Psychometric Laboratory, 1970–present (Psychometrics, statistics, research design, psychoanalytic psychology)

Primary Activities

Developing the philosophical and mathematical foundations and methods necessary to construct practical, objective measurement, especially in the social and health sciences (inferential stability, conjoint additivity, composition analysis). Designing, applying, teaching and publishing better methods for observing, measuring and verifying the measurement of educational, psychological and physical functioning.

Collaborations

School Improvement: For educational test construction, curriculum validation, item function and student performance quality control, standard setting, school assessment, program evaluation and the study of individual development.

Chicago Center for School Improvement
Chicago Public Schools
Consortium on Chicago School Research
Glen Ellyn Consolidated School District 89
Hebrew University, Department of Sociology, Jerusalem
Illinois State Board of Education, field
Kuwait University, College of Arts, Kuwait
Ministry of Education, Research Branch, Singapore
Minneapolis School Board
Nanyang Tech.Univ.Centre for Applied Res. in Education, Singapore
Ngee Ann Polytechnic, Singapore
NorthWest Evaluation Association, Portland
Portland Public Schools
University Illinois at Chicago, Department of Educational Psychology
University of Toledo, College of Education

Educational Associations: For annual and semi-annual presentations and exchanges of new methods and applications of educational and psychological measurement.

American Educational Research Association
AERA Rasch Measurement SIG
Chicago Objective Measurement Education Table
National Council on Measurement in Education
NorthWest Evaluation Association
Michigan Educational Research Association
Midwestern Educational Research Association
Midwest Objective Measurement Seminars
International Objective Measurement Workshops
International Outcome Measurement Conference

Professional Certification: Construction, validation and standard setting of examinations used to certify professional competence.

American Board of Neurological Surgeons
American Board of Neuroscience Nursing
American Board of Orthopedic Surgery
Moss Rehabilitation Research Institute, Philadelphia
Rush University, Division of Psychosocial Oncology
State Univ. New York at Buffalo, Dept. Rehabilitation Medicine
Uniform Data System for Medical Rehabilitation, Buffalo
University of Chicago, Department of Pediatrics
Univ. Illinois at Chicago, Department of Occupational Therapy
Univ. Illinois at Chicago, Department of Physical Therapy
Univ. Illinois at Chicago, School of Public Health
University of Denver, College of Education
University of Extremadura, Faculty of Economics, Spain
U.S. Department of Health Policy and Administration

Editorial

Boards	Educational and Psychological Measurement
	Education Research and Perspectives
	Educational Research Quarterly
	Journal of Outcome Measurement
	Mid-Western Educational Researcher
	Popular Measurement
	Rasch Measurement Transactions
	Journal of Outcome Measurement
Reviewer	Applied Psychological Measurement
	American Educational Research Association
	Archives of Physical Medicine and Rehabilitation
	British Journal Mathematical and Statistical Psychology
	Educational Evaluation and Policy Analysis
	Journal of the American Medical Association
	Journal of Documentation
	Journal of Educational Measurement
	Journal of Educational Statistics
	Multivariate Behavioral Research
	National Council on Measurement in Education
	National Science Foundation
	Psychological Reports
	Psychometrika

Publications in Psychology

Books

6. Wright B.D. **Attitudes To Emotional Involvement and Professional Development in Residential Child Care**. Chicago: University of Chicago, 1957.
49. Wright B.D., Tuska S. **Student and First Year Teachers' Attitudes Toward Self and Others.** Washington: U.S.Office of Education, 1966.
61. David T.G., Wright B.D. **Learning Environments**. Chicago: University of Chicago Press, 1975.
65. Levinsohn F.H., Wright B.D. **School Desegregation: Shadow and Substance.** Chicago: University of Chicago Press, 1976.

141. Wright B.D., Yonke A. **Hero, Villain, Saint: The Psychology of the Heroic in Myth, Fairytale and Autobiography.** New York: Peter Lang, 1990.
245. Bouchard E., Wright B.D. **Kinesthetic Ventures: Informed by the work of F.M.Alexander, Stanislavski, Peirce & Freud**. Chicago: MESA Press, 1997.

Journal Articles

2. Wright B.D. Emotional factors shaping child-care relationships**. Human Development Bulletin.** Chicago: University of Chicago Committee on Human Development, 1954, 28-34.
3. Bettelheim B., Wright B.D. Staff development in a treatment institution**. American Journal of Orthopsychiatry**, 1955, 25, 705-19.
4. Wright B.D., Bettelheim B. Professional identity and personal rewards in teaching. **Elementary School Journal**, 1957, 57, 297-307.
5. Wright B.D.. Psychology in the classroom. **The School Review**, 1957, 65, 490-92.
8. Harper L., Wright B.D. Dealing with emotional problems in the classroom**. Elementary School Journa**l, 1958, 58, 316-25. Reprinted in J.F.Hogary, J.R.Eichorn (Eds.). **The Exceptional Child**. New York: Holt-Dryden, 1960, 354-67.
9. Wright B.D. On behalf of a personal approach to learning. **Elementary School Journal**, 1958, 58, 365-75.
10. Wright B.D. Psychiatric consultation in a residential treatment institution - the psychologist's view. **American Journal of Orthopsychiatry**, 1958, 28, 276-82.
11. Wright B.D. Some personal motives for teaching. **Chicago Schools Journal**, 1958, 40, 65-74.
13. Wright B.D. Identification and becoming a teacher. **Elementary School Journal**, 1959, 59, 361-73.
14. Wright B.D. What price honors? **Elementary School Journal**, 1959, 59, 436.
15. Wright B.D. Should children teach? **Elementary School Journal**, 1960, 60, 353-69.
16. Wright B.D. Gardner B. The effect of color on black and white pictures. **American Psychologist**, 1960, 15, 453.
17. Wright B.D., Rainwater L. The connotative meanings of color. **American Psychologist**, 1960, 15, 453.
18. Wright B.D., Gardner B. The effect of color in black and white pictures. **Perceptual and Motor Skills**, 1960, 11, 301-04. In Inter-Society Color Council Newsletter, 1967, 187, 14-17.
20. Wright B.D.. Love and hate in the act of teaching**. Elementary School Journal**, 1961, 61, 349-62.
21. Wright B.D., Hess R.D., Tuska S. Identificatory origins of the self among fathers. **American Psychologist**, 1961,16, 379.
22. Wright B.D., Loomis E.A., Meyer L. Some differences between schizophrenic, retarded and normal pre-school boys. **American Psychologist**, 1961,16, 353.

23. Wright B.D., Rainwater L. The effect of color on apparent warmth, weight, size, distance and movement**. American Psychologist**, 1961,16, 437.
24. Wright B.D.. Goals and Values reevaluated. **American Journal of Psychology**, 1961, 74, 310-312.
26. Wright B.D., Rainwater L. The meanings of color. **Journal of General Psychology**, 1962, 67, 89-99. Reprinted in **Inter-Society COLOR Council Newsletter**, 1967, 188.
27. Wright B.D.. The influence of hue, lightness and saturation on apparent warmth and weight. **American Journal of Psychology**, 1962, 75, 232-41.
28. Wright B.D., Loomis E.A., Meyer L. The semantic differential as a diagnostic instrument for distinguishing schizophrenic, retarded, and normal pre-school boys. **American Psychologist**, 1962, 17, 297.
29. Wright B.D., Sherman B. Teachers' self-awareness and their evaluation of childhood authority figures. **American Psychologist**, 1962, 17, 336.
30. Wright B.D., Rainwater L. The effect of color on apparent size, distance, and movement. **American Psychologist**, 1962, 17, 369.
31. Wright B.D., Loomis E.A., Meyer L. Observational Q-sort differences between schizophrenic, retarded and normal pre-school boys. **Child Development**, 1963, 34, 169-85.
32. Wright B.D., Sherman B. Who is the teacher? **Theory Into Practice**, 1963, 2, 67-72.
33. Wright B.D., Sherman B. Teachers' self-awareness and their evaluation of childhood authority figures. **The School Review**, 1963, 71, 79-86.
35. Wright B.D., Tuska S. Interpersonal origins of women's plans to teach. **American Psychologist**, 1964,19, 470.
36. Wright B.D., Tuska S. Interpersonal origins of men's plans to teach. **American Psychologist**, 1964,19, 719.
37. Wright B.D., Tuska S. The nature and origin of femininity among women. **American Psychologist**, 1964,19, 724.
38. Wright B.D., Tuska S. The price of permissiveness. **Elementary School Journal**, 1965, 65, 179-83. Reprinted in **Education Today**, July 1965.
39. Wright B.D., Tuska S. How does childhood make a teacher? **Elementary School Journal**, 1965, 65, 235-45. Reprinted in Erickson D.A., **Educational Organization and Administration**. Berkeley: McCutchan, 1977, 372-384.
40. Wright B.D., Tuska S. Review of Winch, Robert F., Identification and its Familial Determinants. Indianapolis: Bobbs Merrill, 1962. **American Journal of Sociology**, 1965, 70, 499-501.
41. Wright B.D., Tuska S. Feminine and masculine components in the identity of women. **Women's Education**, 1965, 4, 5-6.
42. Wright B.D., Tuska S. Postscript on permissiveness. **Elementary School Journal**, 1965, 65, 393-94.
43. Wright B.D., Sherman B. Love and mastery in the child's image of the teacher. **The School Review**, 1965, 75, 89-101.
44. Wright B.D., Tuska S. The influence of institution on changes in self-conception during teacher training**. Proceedings 73rd Annual Convention of**

the American Psychological Association, 1965, 299-300. Reprinted in **American Psychologist**, 1965, 20, 466.

45. Wright B.D., Tuska S. The influence of a teacher model on self-conception during teacher training and experience**. Proceedings 73rd Annual Convention of the American Psychological Associatio**n, 1965, 20, 466.
46. Wright B.D., Tuska S. Childhood influences and the teaching career**. Education Diges**t, 1965, 30, 15-18.
47. Wright B.D. Why do we keep bad images of teachers? **Elementary School Journal**, 1965, 66, 66-67.
48. Wright B.D., Tuska S. The nature and origins of feeling feminine. **British Journal of Social and Clinical Psychology**, 1966, 5, 140-49.
50. Wright B.D., Tuska S. The childhood romance theory of teacher development. **School Review**, 1967, 75, 123-54.
52. Wright B.D. What a school is for. In A. Nielsen, **Lust for Learning**. Skyum, Denmark: New Experimental College Press, 1968, 11-15.
53. Wright B.D. The conflict of love and order. In A. Nielsen, **Lust for Learning**. Skyum, Denmark: New Experimental College Press, 1968, 65-68.
54. Wright B.D. Bad images of good teachers. In A. Nielsen, **Lust for Learning**. Skyum, Denmark: New Experimental College Press, 1968, 249-51.
55. Wright B.D., Tuska S. From dream to life in the psychology of becoming a teacher. **School Review**, 1968, 76, 253-93.
56. Wright B.D., Tuska S. Career dreams of teachers. **Transactions**, 1968, 6, 42-46.
72. Wright B.D. Our reasons for teaching. **Theory Into Practice**, 1977, 16, 225-230.

Publications on Measurement

Books

73. Wright B.D., Mead R.J**. Measurement Models in the Definition and Application of Social Science Variables**. Arlington: U.S.Army Research Institute, 1977.
76. Wright B.D., Stone M.H. **Best Test Design: Rasch Measurement**. Chicago: MESA Press, 1979.
89. Wright B.D., Masters G.N. **Rating Scale Analysis: Rasch Measurement**. Chicago: MESA Press, 1982.
    Wright B.D., Mayers P. **Conversational Statistics for Education and Psychology**. NewYork:McGraw-Hill, 1984.
199. Wright B.D., Stone M.H. **Rasch Measurement Primers**. Wilmington DE: JASTAK, 1992.
    Wright B.D. , Stone M.H. **Measurement Essentials**. Wilmington, DE: Wide Range Inc, 1999.

Monographs

7. Wright B.D. **A Simple Method for Factor Analyzing Two-Way Data**. Chicago: Social Research Inc, 1957.
59. Wright B.D., Douglas G.A. **Best Test Design and Self-tailored Testing**. Research Memorandum No.19, MESA Psychometric Laboratory, Education Department, University Chicago, 1975.
60. Wright B.D., Douglas G.A. **Better Procedures for Sample-free Item Analysis**. Research Memorandum No.20, MESA Psychometric Laboratory, Education Department, University Chicago, 1975.
62. Wright B.D., Mead R.J., Draba R.E**. Detecting and Correcting Test Item Bias with a Logistic Response Model.** Research Memorandum No.22, MESA Psychometric Laboratory, Education Department, University Chicago, 1976.
71. Wright B.D., Bell S.R**. Verifying Unconditional Estimation for Rasch Item Analysis with Simulated Data**. Research Memorandum No.26, MESA Psychometric Laboratory, Education Department, University Chicago, 1977.
78. Wright B.D., Masters G.N. **The Measurement of Knowledge and Attitude**. Research Memorandum No.30, MESA Psychometric Laboratory, Education Department, University Chicago, 1980.
79. Wright B.D., Bell S.R. **Fair and Useful Testing with Item Banks**. Research Memorandum No.32, MESA Psychometric Laboratory, Education Department, University Chicago, 1980.
82. Masters G.N., Wright B.D. **A Model for Partial Credit Scoring**. Research Memorandum No.31, MESA Psychometric Laboratory, Education Department, University Chicago, 1981.
83. Grosse M.E., Wright B.D. **Patient Management Problem Studies: A Technical Report**. Philadelphia, PA: National Board of Medical Examiners, 1981.
93. Wright B.D**. Fundamental Measurement in Social Science and Education**. Research Memorandum No.33, MESA Psychometric Laboratory, Education Department, University Chicago, 1983.
110. Wright B.D., Grosse M.E., Mead R.J. **A Study of Rasch Estimation and Fit Statistics**. Philadelphia, PA: National Board of Medical Examiners, 1986.
112. Douglas G.A., Wright B.D. **The Two Category Model for Objective Measurement**. Research Memorandum No.34, MESA Psychometric Laboratory, Education Department, University Chicago, 1986.
113. Wright B.D., Douglas G.A. **The Rating Scale Model for Objective Measurement**. Research Memorandum No.35, MESA Psychometric Laboratory, Education Department, University Chicago, 1986.
114. Wright B.D., Lunz M.E. **Standards Combining Expert Judgement, Mastery Level and Statistical Confidence**. Research Memorandum No.37, MESA Psychometric Laboratory, Education Department, University Chicago, 1987.
115. Wright B.D. **Bayes' Answer to Perfection**. Research Memorandum No.38, MESA Psychometric Laboratory, Education Department, University Chicago, 1987.
116. Linacre J.M., Wright B.D. **Item Bias: Mantel-Haenszel and the Rasch Model**. Research Memorandum No.39, MESA Psychometric Laboratory, Education Department, University Chicago, 1987.

126. Grosse M.E., Wright B.D. **Fit to the Rasch Model for Client Examinations**. Philadelphia, PA: National Board of Medical Examiners, 1988.
128. Wright B.D., Linacre J.M. **Rasch Measurement of D.O.T. Process Skills Assessment**. Chicago: University of Illinois, Department of Occupational Therapy, 1988.
165. Linacre J.M., Heinemann A.W., Wright B.D., Granger C.V., Hamilton B.B. **The Functional Independence Measure as a measure of disability.** Rehabilitation Services Evaluation Unit Research Report 91-01. Chicago: Rehabilitation Institute Chicago, 1991
167. Heinemann A.W., Linacre J.M., Wright B.D., Granger C.V. **Relationships between impairment and physical disability as measured by the Functional Independence Measure**. Rehabilitation Services Evaluation Unit Research Report 91-02. Chicago: Rehabilitation Institute of Chicago, 1991.

Computer Programs

A series of FORTRAN programs to implement Rasch's new measurement models beginning with**: RASCH** 1964, **BIGPAR**: For Rating Scales 1965 and **RASCAL:** For General Distribution 1970.

58. Wright B.D., Mead R.J. **CALFIT: Sample-Free Item Calibration with a Rasch Measurement Model.** Research Memorandum No.18, MESA Psychometric Laboratory, Education Department, University Chicago, 1975.
70. Wright B.D., Mead R.J. **BICAL: Calibrating Rating Scales with the Rasch Model.** Research Memorandum No.23, MESA Psychometric Laboratory, Education Department, University Chicago, 1977.
135. Wright B.D., Linacre J.M., Schulz E.M**. BIGSCALE: Rasch Analysis Computer Program.** Chicago: MESA Press, 1989.
140. Wright B.D., Schulz E.M. **MFORMS: Rasch Program for One-Step Item Banking of Dichotomous and Partial Credit Data from Multiple Forms**. Chicago: MESA Press, 1990.
164. Wright B.D., Linacre J.M. **BIGSTEPS: Rasch Computer Program for All Two Facet Problems**. Chicago: MESA Press, 1991-96.
170. Linacre J.M., Wright B.D. **FACETS: Many-Faceted Rasch Analysis**. Chicago: MESA Press, 1992-2001.
170. Linacre J.M., Wright B.D. **WINSTEPS: Rasch Analysis**. Chicago: MESA Press, 1996-2001.

Tests

12. Wright B.D. **A Semantic Differential and How to Use It**. Chicago: Social Research Inc, 1958.
64. Gardner B., Stone M.H., Wright B.D**. Observation, Measurement, Analysis Self-Concept Scale**. Chicago: Social Research Inc, 1976.
81. Stone M.H., Wright B.D. **Knox's Cube Test**. Chicago: Stoelting, 1980-96.

Journal Articles and Chapters

1. Townes C.H., Merritt F.R., Wright B.D. The pure rotational spectrum of ICL. **Physical Review**, 1948, 73, 1334-37.
19. Wright B.D., Evitts M. Direct factor analysis in sociometry. **Sociometry**, 1961, 24, 82-98.
25. Wright B.D.. Statistical Procedures, In S.Lichter, E.Rapien, F.Seiberg, M.Sklansky, **The Drop-Outs**. New York: Free Press, 1962, 270-82.
34. **Wright** B.D., Evitts S. Multiple regression in the explanation of social structure. **Journal of Social Psychology**, 1963, 61, 87-98.
51. Wright B.D.. Sample-free test calibration and person measurement. In **Proceedings 1967 Invitational Conference on Testing.** Princeton: Educational Testing Service, 1968, 85-101.
57. Wright B.D., Panchapakesan N. A procedure for sample-free item analysis. **Educational and Psychological Measurement**, 1969, 29, 23-48.
66. Wright B.D., Douglas G.A. Best procedures for sample-free item analysis. **Applied Psychological Measurement**, 1977, 1, 281-295.
67. Wright B.D.. Solving measurement problems with the Rasch model. **Journal of Educational Measurement**, 1977, 14, 97-116.
68. Wright B.D.. Misunderstanding the Rasch model. **Journal of Educational Measurement**, 1977, 14, 219-226.
69. Wright B.D., Douglas G.A. Conditional versus unconditional procedures for sample-free item analysis. **Educational and Psychological Measurement**, 1977, 37, 573-586.
75. Perline R., Wright B.D., Wainer H. The Rasch model as additive conjoint measurement. **Applied Psychological Measurement**, 1979, 3, 237-255.
77. Wright B.D. Foreword and Afterword. In G. Rasch, **Probabilistic Models**. Chicago: University Chicago Press, 1980.
80. Wainer H., Wright B.D. Robust estimation of ability in the Rasch model. **Psychometrika**, 1980, 45, 373-391.
84. Schulz E.M., Lambert R.W., Wright B.D., Becker S.W. An overview of the blind rehabilitation process**. Proceedings Fourth Annual Conference on Rehabilitative Engineering**, 1981, 1, 94-96.
85. Masters G.N., Wright B.D. Defining a Fear-of-Crime variable: a comparison of two Rasch models. **Education Research and Perspectives**, 1982, 9, 18-31.
86. **Lambert** R.W., Becker S.W., Courington S.M., Wright B.D. Evaluating the rehabilitation process: an example with the blind. **International Journal of Rehabilitation Research**, 1982, 5, 487-498.
87. Schulz E.M., Wright B.D, Lambert R.W., Becker S.W., Ludlow L.H. A measure of activity capacity in blind rehabilitation. **Proceedings Fifth Annual Conference on Rehabilitation Engineering**, 1982, 2, 112-113.
88. Ludlow L. H., Wright B.D., Lambert R.L., Becker S.W. The measurement of attitudes toward blindness. **Proceedings 1982 Conference Rehabilitation Society of North America**. ERIC Document: ED 222-523.

90. Courington S.M., Lambert R.W., Wright B.D., Becker S.W., Ludlow L.H. The measurement of attitudes toward blindness and its importance for rehabilitation. **International Journal of Rehabilitation Research**, 1983, 6, 67-72.
91. Stone M.H., Wright B.D. Measuring attending behavior and short-term memory with Knox's cube test. **Educational and Psychological Measurement**, 1983, 43, 803-815.
92. Wright B.D., Stone M.H. Review of The British Ability Scales. An-09032764. Buros Institute Database Bibliographic Retrieval Services**. Ninth Mental Measurements Yearbook.** Lincoln Neb: Buros Institute, 1983.
94. Schulz E.M., Wright B.D., Lambert R.W., Becker S.W. Measuring change in activity capacity in blind rehabilitation**. Proceedings Sixth Annual Conference on Rehabilitation Engineers**, 1983, 3, 327-329.
95. Ludlow L.H., Wright B.D., Lambert R.W., Becker S.W. Measuring change with the rating scale model**. Proceedings 1983 Conference American Educational Research Association**. ERIC Document: ED-228-324.
96. Grosse,M.E., Wright B.D., Schumacher C.F. Equating examinations with the Rasch model. In Lloyd, Langsley, **Evaluating the Skills of Medical Specialists**. Chicago: American Board of Medical Specialties, 1983, 273-282.
97. Wright B.D. Rasch measurement models. **International Encyclopedia of Education.** Oxford: Pergamon, 1984.
98. Wright,B.D.. Essay review of "The Improvement of Measurement in Education and Psychology." **Australian Journal of Education**, 1984, No. 2.
100. Wright B.D.. Despair and hope for educational measurement**. Contemporary Education Review**, 1984, 1, 281-288.
101. Wright B.D., Bell S.R. Item banks: what, why, how**. Journal of Educational Measurement**, 1984, 21, 331-345.
102. Masters G. N., Wright B.D. The essential process in a family of measurement models. **Psychometrika**, 1984, 49, 529-544
103. Grosse M.E., Wright B.D. Validity and reliability of true-false tests**. Educational and Psychological Measurement**, 1985, 45, 1-14.
104. Wright B.D. Additivity in psychological measurement. In Edw. Roskam, **Measurement and Personality Assessment**. Amsterdam: North-Holland, 1985, 101-112.
108. Lambert R.W., Wright B.D. Measuring attitudes with unidimensional scaling and factor analysis. **Journal of Rehabilitation Research**, 1985, 8, 415-424.
109. Schulz E.M. , Wright B.D., Lambert R.W., Becker S.W., Bezruczko N. An assessment of the needs of rehabilitated blind veterans. **Journal of Visual Impairment and Blindness**, 1985, 79, 301-305.
111. Grosse M.E., Wright B.D. Setting and maintaining certification standards. **Evaluation and the Health Professions**, 1986, 9, 267-285.
119. Grosse M.E., Wright B.D. Psychometric characteristics of scores on a patient management problem test. **Educational and Psychological Measurement**, 1988, 48, 297-305.
121. Julian E.R., Wright B.D. Using computerized patient simulations to measure the clinical competence of physicians. **Applied Measurement in Education**, 1988, 1, 299-318.

122. Wright B.D. Rasch measurement models. In J. P. Keeves (Ed.), **Educational Research, Methodology, and Measurement: An International Handbook**. London: Pergamon, 1988, 286-292.
123. Wright B.D.. The efficacy of unconditional maximum likelihood bias correction. **Applied Psychological Measurement**, 1988, 12, 315-318.
133. Wright B.D., Linacre,J.M. Observations are always ordinal: Measures, however, must be interval. **Archives of Physical Medicine and Rehabilitation**, 1989, 70, 857-860.
148. Lunz M. E., Wright B.D., Linacre J.M. Measuring the impact of judge severity on examination scores. **Applied Measurement in Education**, 1991, 4.
149. **Wright** B.D. The International Objective Measurement Workshop: Past and Future. In M. Wilson (Ed.) **Objective Measurement: Theory into Practice I.** Norwood NJ: ABLEX, 1991, 9-28.
151. Julian E.R., Wright B.D. Distinguishing Between Shared and Unique Employee Needs. In M. Wilson (Ed.) **Objective Measurement: Theory into Practice I**, Norwood NJ: ABLEX, 1991, 97-108.
150. Schulz E.M., Perlman C., Rice W.K., Wright B.D. Vertically Equating Reading Tests. In M. Wilson (Ed.) **Objective Measurement: Theory into Practice I**, Norwood NJ: ABLEX, 1991, 138-156.
Heinemann A.W., Linacre J.M., Wright B.D., Granger C.V., Hamilton B.B. Prediction of Rehabilitation Outcomes following Spinal Cord Injury with the Functional Independence Measure. **Journal of the American Paraplegia Society**, 1992.
181. Lunz M.E., Bergstrom B.A., Wright B.D. The Effect of Review on Student Ability and Test Efficiency for Computerized Adaptive Tests. **Applied Psychological Measurement**, 1992, 16, 1, 33-40.
176. Granger C.V., Hamilton B.B., Linacre J.M., Heinemann A.W., Wright B.D. Performance Profiles of the Functional Independence Measure. **American Journal of Physical Medicine and Rehabilitation**, 1993, 72, 84-89.
177. Heinemann A.W., Linacre J.M., Wright B.D., Granger C.V. Relationships between Impairment and Physical Disability as Measured by the Functional Independence Measure. **Archives of Physical Medicine and Rehabilitation**, 1993, 74, 566-573.
180. Wright B.D., Linacre J.M. Heinemann A.W. Measuring Functional Status in Rehabilitation. In C.V.Granger, G.E.Gresham (Eds.). **Physical Medicine and Rehabilitation Clinics of North America: New Developments in Functional Assessment**, 1993, 4, 475-492.
219. Granger C.V., Wright B.D. Looking Ahead to the Use of Functional Assessment in Ambulatory Physiatric and Primary Care: The Functional Assessment Screening Questionnaire. **Physical Medicine and Rehabilitation Clinics of North America**, 1993, 4, 1-11.
173. Lunz M.E., Bergstrom B.A., Wright B.D. Reliability of Alternate Computer Adaptive Tests. In M.Wilson (Ed.) **Objective Measurement: Theory into Practice II**. Norwood NJ: ABLEX, 1994, 115-121.

211. Adams R., Wright B.D. When does misfit make a difference? In Mark Wilson (Ed.) **Objective Measurement: Theory Into Practice II**. New Jersey: ABLEX, 1994. Pages 244-270.
184. Wright B.D. Composition Analysis. **Mid-Western Educational Researcher**, 1994, 7, 29-38.
186. Fisher W.P., Wright B.D. (Ed.s) Applications of Probabilistic Conjoint Measurement. Special Issue. **International Journal Educational Research**, 1994, 21, 557-664.
192. Heinemann A.W., Linacre J.M., Wright B.D., Hamilton B.B., Granger C.V. Measurement Characteristics of the Functional Independence Measure**. Topics in Stroke Rehabilitation**, 1994, 1, 1-15.
193. Fisher A.G., Bryze K.A., Granger C.V., Haley S.M., Hamilton B.B., Heinimann A.W., Jalayerian J.K.,Linacre J.M., Ludlow L.H., McCabe M.A., Wright B.D. Applications of Rasch Analysis to the Development of Functional Assessments**. International Journal Educational Research**, 1994, 21, 579-594.
194. Linacre J.M., Wright B.D. Constructing Linear Measures from Counts of Qualitative Observations. **Fourth International Conference on Bibliometrics, Informetrics and Scientometrics**, Berlin, Germany. ERIC: TM020794, September, 1994.
200. Linacre J.M., Heinemann A.W., Wright B.D., Granger C.V. Hamilton B.B. The Structure and Stability of the Functional Independence Measure**. Archives Physical Medicine and Rehabilitation**, 1994, 75, 127-132.
202. Heinemann A.W., Linacre J.M., Wright B.D., Hamilton B.B., Granger C.V. Prediction of Rehabilitation Outcomes with Disability Measures. **Archives Physical Medicine and Rehabilitation**, 1994, 75, 133-143.
210. Lunz M.E., Stahl, J.A., Wright B.D. Interjudge Reliability and Decision Reproducibility. **Educational and Psychological Measurement**, 1994, 54, 913-925.
212. Fisher W.P., Wright B.D. Introduction to Probabilistic Conjoint Measurement Theory and Applications. **International Journal Educational Research**, 1994, 21, 559-568.
213. Smith R., Julian E., Lunz M., Stahl J. Schulz M., Wright B.D. Applications of Conjoint Measurement in Professional Certification Programs. **International Journal of Educational Research**, 1994, 21, 653-664.
226. Heinemann A.W., Hamilton B.B., Linacre J.M., Wright B.D., Granger C.V. Functional Gains and Therapeutic Intensity during Rehabilitation. **American Journal Physical Medicine and Rehabilitation**, 1995, 74, 315-326.
Schulz E.M., Perlman C.,Rice W.K., Wright B.D. An Empirical Comparison of Rasch and Mantel-Haenszel Procedures for Assessing Differential Item Functioning. In G.Engelhard, M.Wilson (Ed.s) **Objective Measurement: Theory into Practice III**. Norwood, NJ: Ablex, 1996, 65-82.
182. Lunz M.E., Stahl J.A., Wright B.D. The Invariance of Judge Severity Calibrations. In G.Engelhard, M.Wilson (Ed.s) **Objective Measurement: Theory into Practice III**. Norwood NJ: Ablex, 1996, 99-112.

201. Wright, B.D. Composition Analysis. In George Engelhard, Mark Wilson (Eds.) **Objective Measurement: Theory into Practice III**. Norwood, NJ: Ablex, 1996, 241-264.
223. Wright B.D. Comparing Rasch Measurement and Factor Analysis. **Structural Equation Modeling**, 1996, 3, 3-24.
227. Cella D.F., Lloyd S.R., Wright B.D. Cross-Cultural Instrument Equating. In B.Spilker (Ed) **Quality of Life and Pharmacoeconomics in Clinical Trials**. New York: Lippencott-Raven, 1996, 73, 707-715.
228. Lunz M.E., Wright B.D., Stahl J.A. Applications of the Multi-Facet Rasch Model to Medical Certification Performance Examinations. In J.Rost (Ed) **Proceedings IPN Sankelmark Symposium on Applications of Latent Trait and Latent Class Models in the Social Sciences.** Akademie Sankelmark Germany, 1996.
239. Grimby G., Andren E., Holmgren E., Wright B.D., Linacre J.M., Sundh V. Structure of a Combination of Functional Independence Measure and Instrumental Activity Measure Items in Community-Living Persons: A Study of Individuals With Cerebral Palsy and Spina Bifida. **Archives of Physical Medical Rehabilitation**, 77, 1996, 1109-1114.
240. Kindlon D.J., Wright B.D., Raudenbush S.W., Felton E. The Measurement of Children's Exposure to Violence: A Rasch Analysis. **International Journal of Methods in Psychiatric Research**, 6, 1996, 187-194.
248. Nordenskiold U., Grimby G., Hedberg M., Wright B.D., Linacre J.M. The Structure of an Instrument for Assessment of the Effect of Assistive Devices and Altered Work Methods in Women with Rheumatoid Arthritis. **Arthritis Care Research**, 9, 5, 1996, 358-367.
191. Heinemann A.W., Kirk P., Hastie B.A., Semik P., Hamilton B.B., Linacre J.M., Wright B.D., Granger C.V. Relationships between Disability and Nursing Effort during Medical Rehabilitation for Patients with Traumatic Brain and Spinal Cord Injury. **Archives of Physical Medicine Rehabilitation**, 78, 1997, 143-149.
    Wright BD, Linacre JM, Smith RM, et al. FIM measurement properties and rasch model details **Scandinavian Journal of Rehabilitation Medicine**, 4, 1997, 267-270
229. Lunz M.E., Wright B.D. Latent Trait Models for Performance Evaluations. In J.Rost and R. Langehiene (Eds). **Applications of Latent Trait and Latent Class Models in Social Sciences**. New York: Waxmann, 1997, 80-88.
241. Masters G.N., Wright B.D. The Partial Credit Model. In W.J. Linden, R.K. Hambleton. (Eds) **Handbook of Modern Item Response Theory**. New York: Springer-Verlag. 1997, 101-122.
244. Wright B.D. Fundamental Measurement for Outcome Evaluation. In R. Smith (Ed) **Outcome Measurement. PM&R: State of the Art Reviews**. Philadelphia: Hanley & Belfus. 1997, 261-288.
    M.E.Segal, A.W.Heinemann, R.R.Schall, Wright B.D. Rasch Analysis of a Brief Physical Ability Scale for Long-Term Outcomes of Stroke. In R.Smith (Ed) **Outcome Measurement. PM&R: State of the Art Reviews**. Philadelphia: Hanley & Belfus. 1997, 385-396.

246. Bookstein A, Wright B.D. Ambiguity in measurement. **Scientometrics**, 40, 3, 1997, 369-384.
Du Y., Wright B.D. Effects of Item Characteristics in a Large-scale Direct Writing Assessment. In M.Wilson, G.Engelhard, Draney K. (Ed.s) **Objective Measurement: Theory into Practice IV**. Norwood NJ: Ablex, 1997, 1-24.
Wright B.D. Rasch Factor Analysis. In M.Wilson, G.Engelhard, Draney K. (Ed.s) **Objective Measurement: Theory into Practice IV**. Norwood NJ: Ablex, 1997, 113-138.
247. Wright B.D. How To Measure Outcomes. **Rehabilitation Outlook**, 3, 1, 1998, 6-9.
Grimby G., Andren E., Daving Y., and Wright B.D. (1998) Dependence and perceived difficulty in daily activities in community-living stroke survivors 2 years after stroke - a study of instrumental structures. **Stroke**, 29, 9, 1843-1849.
249. Prieto L., Alonso J., Lamarca R., Wright B.D. Rasch Measurement for Reducing the Items of the Nottingham Health Profile. **Journal of Outcome Measurement**, 2, 4, 1998, 285-301.
Wright B.D. A History of Social Science Measurement. **Educational Measurement: Issues and Practice.** 16, 4, 1998, 33-52.
Ryser L, Wright BD, Aeschlimann A, et al. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis **Arthrit Care Res**, 12, 5, 1999, 331-335.
Wright B.D. Fundamental Measurement for Social Science Research. In S.Embretson, S.Hershberger. (Eds) **The New Rules of Measurement: What Every Psychologist and Educator Should Know**. Hillsdale, NJ: Lawrence Erlbaum Ass. 1999, 65-104.
Chang C.H., Wright B.D., Cella D., Hayes R.D. Re-examination of Physical and Mental Health as Measured by the RAND-36. **Association for Health Services Research**, 1999, 12.
Cella D., Chang C.H., Wright B.D., Von Roenn J., Skeel R. Defining High-Order Dimensions of Self-Reported Health. **Quality of Life Research**, 8, 7, 1999, 598.
Wright B.D., Mok M. Understanding Rasch Measurement: Rasch Models Overview. **Journal of Applied Measurement**, 1, 1 2000, 83-106.
Wright B.D., Perkins K.,Dorsey J. Rasch Measurement Instead of Regression. **Multiple Linear Regression Viewpoints**, 26, 2, 2000, 36-41.
Wright B.D. Multiple Regression with WINSTEPS. **Multiple Linear Regression Viewpoints**, 26, 2, 2000, 42-45.

Periodicals and Presentations

63. Wright B.D., Mead, R.J. Fit analysis of a reading comprehension test. San Francisco: **AERA,** 1976.
105. Iwamoto C., Kelley P., Wright B.D. The calculation of cutting scores from a bank of Rasch item difficulties. **Third International Objective Measurement Workshop**. Chicago: MESA Press, 1985.

106. Erviti V.F., Pappas J.S., Wright B.D. The use of rating scales in the education and evaluation of medical students. **Third International Objective Measurement Workshop**. Chicago: MESA Press, 1985.
107. Grosse M.E., Wright B.D. How Rasch measurement can help with professional certification. **Third International Objective Measurement Workshop**. Chicago: MESA Press, 1985.
117. Wright B.D., Linacre J.M. Rasch model derived from objectivity. **Rasch Measurement Transactions**, 1, 1, 1987, 5.
118. Wright B.D. Some comments about guessing. **Rasch Measurement Transactions**, 1, 2, 1987, 6.
120. Wright B.D. How interaction denies objectivity. **Rasch Measurement Transactions**, 1, 2, 1988, 12.
124. Wright B.D. Rasch model from Thurstone's scaling requirements. **Rasch Measurement Transactions**, 2, 1, 1988, 13-14.
125. Wright B.D. Rasch model from Campbell concatenation. **Rasch Measurement Transactions**, 2, 1, 1988, 16.
Wright B.D. Practical adaptive testing. **Rasch Measurement Transactions**, 2, 2, 1988, 24.
Review of cooperation between B D Wright and G Rasch. **Rasch Measurement Transactions**, 2, 2, 1988, 19 .
Wright B.D. Georg Rasch and measurement. **Rasch Measurement Transactions** 2, 3, 1988, 25.
Schulz E.M., Wright B.D. One-step vertical test equating. San Francisco: **AERA**, 1989.
130. Schulz E.M., Wright B.D. An empirical comparison of Rasch and Mantel-Haenszel procedures. San Francisco: **NCME**, 1989.
Linacre J.M., Wright B.D. Length of a logit. **Rasch Measurement Transactions**, 3, 2, 1989. 54.
136. Linacre J.M., Wright B.D. Equivalence of Rasch PROX and Mantel-Haenzel. **Rasch Measurement Transactions**, 3, 2, 1989, 57.
Wright B.D. Rasch model from counting right answers. **Rasch Measurement Transactions**, 3, 2, 1989, 62.
134. Wright B.D., Linacre, J.M. The differences between scores and measures. **Rasch Measurement Transactions**, 3, 3, 1989, 63.
138. Lunz M.E., Stahl J.A., Wright B.D. Variations among examiners and protocols on oral examinations. San Francisco: **AERA**, 1989.
139. Lunz M.E., Stahl J.A., Wright B.D. Equating practical examinations. San Francisco: **NCME**, 1989.
Wright B.D., Masters G.N. Outfit and Infit. **Rasch Measurement Transactions**, 3, 4, 84 .
MESA Psychometric Laboratory. Wright B.D. **Rasch Measurement Transactions**, 3, 4, 87, 1989.
143. Schulz E.M., Shen L.S., Wright B.D. Constructing an equal-interval scale for growth in Reading Achievement. Boston: **AERA**, 1990.

144. Lunz M.E., Wright B.D. Setting criterion standards from benchmark performances. Boston: **AERA**, 1990.
145. Lunz M.E., Bergstrom B., Gershon R., Wright B.D. Test-retest consistency of computer adaptive tests. Boston: **NCME**, 1990.
Wright B.D. What is Information? **Rasch Measurement Transactions**, 4, 2, 1990, 109.
Wright B.D. Glossary of misleading terms. Wright BD. **Rasch Measurement Transactions**, 4, 3, 1990, 116.
147. Bergstrom B.A., Lunz M.E., Wright B.D. The stability of Rasch pencil and paper item calibrations on computer adaptive tests. Chicago: **Midwest Objective Measurement Seminar**, June 1990.
152. Wright B.D., Stone M.H. The Idea of a Variable. **Measurement Primer** 3. Wilmington DE: JASTAK, 1991.
153. Wright B.D., Stone M.H. Deducing the Measurement Model. **Measurement Primer** 4. Wilmington DE: JASTAK, 1991.
154. Wright B.D., Stone M.H. Fit Analysis. **Measurement Primer** 8. Wilmington DE: JASTAK, 1991.
155. Wright B.D., Stone M.H. Identifying Item Bias. **Measurement Primer** 9. Wilmington DE: JASTAK, 1991.
156. Wright B.D., Stone M.H. Control Lines for Item Plots. **Measurement Primer** 10. Wilmington DE: JASTAK, 1991.
157. Wright B.D., Stone M.H. Building Scholastic Variables. **Measurement Primer** 14. Wilmington DE: JASTAK, 1991.
158. Wright B.D., Stone M.H. Estimating Item Calibrations and Person Measures. **Measurement Primer** 18. Wilmington DE: JASTAK, 1991.
159. Wright B.D., Stone M.H. Calibration by Hand. **Measurement Primer** 19. Wilmington DE: JASTAK, 1991.
160. Wright B.D., Stone M.H. Information and Misfit Analysis. **Measurement Primer** 20. Wilmington DE: JASTAK, 1991.
161. Wright B.D., Stone M.H. Separation Statistics. **Measurement Primer** 21. Wilmington DE: JASTAK, 1991.
162. Wright B.D., Stone M.H. Reliability. **Measurement Primer** 22. Wilmington DE: JASTAK, 1991.
163. Wright B.D., Stone M.H. Validity. **Measurement Primer** 23. Wilmington DE: JASTAK, 1991.
166. Lunz M.E., Stahl J.A., Wright B.D. Invariance of Judge Severity Calibrations. Chicago: **AERA**, 1991.
168. Linacre J.M., Heinemann A.W., Wright B.D. Rating Scale Analysis of the Functional Independence Measure. Washington: **American Congress of Rehabilitation Medicine**, 1991.
169. Wright B.D. Factor Analysis versus Rasch. **Rasch Measurement Transactions**. 5, 1, 1991, 134.
171. Wright B.D. Diagnosing Misfit. **Rasch Measurement Transactions**. 5, 2, 1991, 156.

175. Wright B.D. Errors, Variances Correlations. **Rasch Measurement Transactions**. 5, 2, 1991, 147.
Wright B.D. Georg Rasch's BPP. **Rasch Measurement Transactions**. 5, 3, 1991, 169.
188. Wright B.D. Scores, Reliability, Assumption. **Rasch Measurement Transactions**. 5, 3, 1991,
Wright B.D. New Standards Project. Resnick L, Wright BD. **Rasch Measurement Transactions**, 5, 3, 1991, 168.
189. Wright B.D. Understanding Confidence Intervals. **Rasch Measurement Transactions**. 5, 4, 1992, 175.
190. Wright B.D. Point-Biserials and Item Fits. **Rasch Measurement Transactions**. 5, 4, 1992, 174.
195. Wright B.D. Rasch versus Birnbaum". **Rasch Measurement Transactions**. 5, 4, 1992, 178-179.
196. Wright B.D. IRT in the 1990's. **Rasch Measurement Transactions**. 6, 1, 1992, 196.
197. Wright B.D. Scores are Not Measures. **Rasch Measurement Transactions**. 6, 1, 1992, 208.
198. Wright B.D. What is "Right" Length? **Rasch Measurement Transactions**. 6, 1, 1992, 205.
203. Wright B.D. Rasch Model from Ratio Scale Counts. **Rasch Measurement Transactions** 6, 2, 1992, 219.
204. Wright B.D. The Rasch Family of Two-Facet Models. **Rasch Measurement Transactions** 6, 2, 1992, 226.
205. Wright B.D., Linacre J.M. Combining Categories. **Rasch Measurement Transactions** 6, 3, 1992, 233.
206. Wright B.D. Anchoring and Standard Errors. **Rasch Measurement Transactions** 6, 4, 1993, 259.
207. Wright B.D. Equitable Test Equating. **Rasch Measurement Transactions** 7, 2, 1993, 298.
208. Wright B.D. Logits. **Rasch Measurement Transactions** 7, 2, 1993, 288.
209. Wright B.D. Thinking with Raw Scores. **Rasch Measurement Transactions** 7, 2, 1993, 299.
214. Wright B.D. Survival Analysis. **Rasch Measurement Transactions** 7, 3, 1993, 307.
215. Wright B.D., Grosse M.E. How to Set Standards. **Rasch Measurement Transactions** 7, 3, 1993, 315.
216. Wright B.D. Data analysis and fit. **Rasch Measurement Transactions**, 7, 4, 1994, 324.
217. Wright B.D. Where Dimensions Come From? **Rasch Measurement Transactions**, 7, 4, 1994, 326.
218. Wright B.D. Rasch Factor Analysis. **Rasch Measurement Transactions**, 8, 1, 1994, 348.
220. Wright B.D. Foundations of Inference. **Rasch Measurement Transactions**, 8, 1, 1994, 346.

221. Stenner J., Wright B.D., Linacre J.M. From P-values to Logits. **Rasch Measurement Transactions**, 8, 1, 1994, 338.
230. Linacre J.M., Wright B.D. Chi-Square Fit Statistics. **Rasch Measurement Transactions**, 8, 2, 1994, 360.
Green K.E., Kluever R.C., Wright B.D. Predicting item difficulties from item characteristics. **Rasch Measurement Transactions**, 8, 2, 1994, 354.
231. Wright B.D. Theory Construction from Empirical Observations. **Rasch Measurement Transactions**, 8, 2, 1994, 362.
232. Wright B.D., Linacre J.M. Reasonable Mean-Square Fits. **Rasch Measurement Transactions**, 8, 3, 1994, 370.
233. Wright B.D. Measuring and Counting. **Rasch Measurement Transactions**, 8, 3, 1994, 371.
234. Wright B.D. Part-Test vs. Whole-Test Measures. **Rasch Measurement Transactions**, 8, 3, 1994, 376.
235. Andrich D., Wright B.D. Rasch Sensitivity and Thurstone Insensitivity to Graded Responses. **Rasch Measurement Transactions**, 8, 3, 1994, 382.
236. Roberts J., Stone M., Wright B.D. Maximizing Rating Scale Information. **Rasch Measurement Transactions**, 8, 3, 1994, 386.
Wright B.D. Unidimensionality coefficient. **Rasch Measurement Transactions**, 8, 3, 1994, 385
Stone M.H., Wright B.D. Maximizing rating scale information.. **Rasch Measurement Transactions**, 8, 3, 1994, 386
237. Wright B.D. Reading in America: Stenner's Lexiles Confirmed! **Rasch Measurement Transactions**, 8, 4, 1995, 387-388.
Wright B.D. Rasch and Wright: the early years. In Linacre, J.M. (Ed) **Rasch Measurement Transactions**, Part 1. Chicago: Mesa Press, 1995, 1-4.
238. Linacre J.M., Wright B.D. How to Assign Item Weights - If you Must. **Rasch Measurement Transactions**, 8, 4, 1995.403
Wright B.D. Problem drinking. **Rasch Measurement Transactions**, 8, 4, 1995, 402.
Wright B.D. 3PL or Rasch? Wright BD. **Rasch Measurement Transactions**, 9, 1, 1995, 408.
Rudner L., Wright B.D. Diagnosing person misfit. **Rasch Measurement Transactions**, 9, 2, 1995, 430.
Wright B.D. Teams, packs and chains. **Rasch Measurement Transactions**, 9, 2, 1995, 432-433.
Linacre J.M., Wright B.D. Measures, correlations and explained variances. **Rasch Measurement Transactions**, 9, 2, 1995, 435.
Wright B.D. Which standard error? **Rasch Measurement Transactions**, 9, 2, 1995, 436-437.
Wright B.D. Majority rule. **Rasch Measurement Transactions**, 9, 3, 1995, 443.
Wright B.D. Sample size again. **Rasch Measurement Transactions**, 9, 4, 1996, 468.

Wright B.D. Reliability and separation. **Rasch Measurement Transactions**, 9, 4, 1996, 472.
Wright B.D. Time 1 to Time 2 comparison. **Rasch Measurement Transactions**, 10, 1, 1996, 478-479.
Wright B.D. Construct problems with descriptive IRT. **Rasch Measurement Transactions**, 10, 1, 1996, 481.
Linacre J.M., Wright B.D. Guttman-style item location maps. **Rasch Measurement Transactions**,10, 2, 1996, 492-493.
Wright B.D. Key events in Rasch measurement history in America, Britain and Australia (1960-1980). **Rasch Measurement Transactions**, 10, 2, 1996, 494-495.
Wright B.D. Pack to Chain to Team? **Rasch Measurement Transactions**, 10, 2, 1996, 501.
Wright B.D. Negative information. **Rasch Measurement Transactions**, 10, 2, 1996, 504.
Wright B.D. Semiotics and scientific method. **Rasch Measurement Transactions**, 11, 1, 1997, 539-540.
Wright B.D. Managing multidimensionality. **Rasch Measurement Transactions**, 11, 1, 1997, 540.
Wright B.D. Stevens revisited. **Rasch Measurement Transactions**, 11, 1, 1997, 552-553.
Wright B.D. Fundamental measurement. **Rasch Measurement Transactions**, 11, 2, 1997, 558.
Wright B.D. The Road to Reason. **Rasch Measurement Transactions**, 11, 4, 1998, 589.
Wright B.D. Interpreting Reliabilities. **Rasch Measurement Transactions**, 11, 4, 1998, 602.
Wright B.D. Two-item testing? **Rasch Measurement Transactions**, 12, 2, 1998, 627-8.
Wright B.D. Who is awarded first prize? **Rasch Measurement Transactions**, 12, 2, 1998, 629.
Wright B.D. Estimating measures for extreme scores. **Rasch Measurement Transactions**, 12, 2, 1998, 632-3.
Wright B.D. Subset fit. **Rasch Measurement Transactions**, 12, 2, 1998, 635.
Wright B.D. Rank-ordered raw scores imply the Rasch model. **Rasch Measurement Transactions**, 12, 2, 1998, 637-8.
Wright B.D. Rasch: The Man Behind the Model. **Popular Measurement**, 1998, 1, 1, 15-22.
Wright B.D. Where Do Dimensions Come From? **Popular Measurement**, 1998, 1, 1, 32.
Wright B.D. What is the "Right" Test Length? **Popular Measurement**, 1998, 1, 1,34.
Wright B.D. Model selection: Rating Scale or Partial Credit? **Rasch Measurement Transactions**, 1999, 12, 3, 641-2.

Barrett P., Wright B.D., Fisher W.P. Jr. Does Rasch Construct Bad Rulers? **Rasch Measurement Transactions**, 1999, 12, 4, 659-660.
Wright B.D., Tuska S.A. Identifying Psychological Variables. **Rasch Measurement Transactions**, 1999, 12, 4, 672.
Wright, B.D. Common Sense for Measurement. **Rasch Measurement Transactions**, 13, 3, 1999, 704-705.
Wright, B.D. Life and Mind. **Rasch Measurement Transactions** 13, 3, 1999, 713-714.
Wright, B.D., Stenner, A.J. One Fish, Two Fish: Rasch Measures Reading Best. **Popular Measurement**, 2, 1, 1999, 34-38.
Wright, B.D., Stenner, A.J. Lexile Perspectives. **Popular Measurement**, 2, 1, 1999, 39-40.
Wright B.D., Stenner, A.J. Using Lexiles. **Popular Measurement**, 2, 1, 1999, 41-42.
Wright B.D., Stenner, A.J. Lexile Perspectives. **Popular Measurement**, 3, 1, 2000, 14-17.
Wright B.D. Rasch Analysis for Surveys. **Popular Measurement**, 3, 1, 2000, 61.
Wright B.D. What's to Learn in Psychometrics. **Popular Measurement**, 3, 1, 2000, 73.
Wright B.D. Three "C's" to Meaning: The Big Picture. **Popular Measurement**, 3, 1, 2000, 74.
Wright B.D. The Road to Reason. **Popular Measurement**, 3, 1, 2000, 75.
Wright B.D. Realizations of Measurement. **Popular Measurement**, 3, 1, 2000, 76.
Wright B.D. Basic Research Methods. **Popular Measurement**, 3, 1, 2000, 77.
Wright B.D. Multiple Regression via Measurement. **Rasch Measurement Transactions**, 14,1,2000,729-731.
Wright B.D. Evolution of Meaning in Practice. **Rasch Measurement Transactions**, 14,1,2000,736-737.
Wright B.D. How to Set Standards. **Rasch Measurement Transactions**, 14, 1, 2000,740-742.
Wright, B.D., Huber, M., O'Neill, T., Linacre, J.M. The Problem of Measure Invariance. **Rasch Measurement Transactions**, 14, 2, 2000,745.
Wright, B.D. Conventional Factor Analysis vs. Rasch Residual Factor Analysis. **Rasch Measurement Transactions** 14, 2, 2000, 753.
Wright, B.D. Rasch Regression: My Recipe. **Rasch Measurement Transactions**, 14, 3, 2000,758-9.
Wright, B.D. Counts or Measures? Which Communicate Best? **Rasch Measurement Transactions**, 14, 4, 2000,784.
Wright, B.D. Separation, Reliability and Skewed Distributions. **Rasch Measurement Transactions**, 14, 4, 2000, 786.

In Process

. Granger C.V., Fiedler R.C., Wright, B.D. The Painfree Measure: Outpatient Physiatric Follow-Along Part II. **American Journal of Physical Medicine and Rehabilitation.**
. Granger C.V., Fiedler R.C., Wright, B.D. The Placid versus Distress Measure: Outpatient Physiatric Follow-Along Part III. **American Journal of Physical Medicine and Rehabilitation.**
. Thomee R., Grimby G., Wright B.D. Rasch Analysis of Visual Analog Scale Measurements Before and After Treatment of Patients with Patellofemoral Pain Syndrome. **Scandinavian Journal of Rehabilitation Medicine.**
. Halper A.S., Cherney L.R., Heinemann A., Semik, P., Wright B.D. Test for Right Hemisphere Communication Problems: Evaluating Psychometric Properties. **American Speech-Language-Hearing Association.**
. Heinemann A.W., Kirk P., Hamilton B.B., Linacre J.M., Wright B.D., Granger C.V. Relationships between Disability and Nursing Effort during Medical Rehabilitation for Patients with Traumatic Brain and Spinal Cord Injury.
. Prieto L., Wright B.D. Rasch Measurement for Reducing the Items of the Nottingham Health Profile. **Journal of Outcome Measurement.**
. Grimby G., Andren E., Daving Y., Wright B.D. Dependence and Perceived Difficulty in Daily Activities in Community-Living Stroke Survivors Two Years after Stroke: A Study of Instrumental Structures. **Stroke**.

In Press

Granger C.V., Fiedler R.C., Wright, B.D. The Painfree Measure: Outpatient Physiatric Follow-Along Part II. **American Journal of Physical Medicine and Rehabilitation**.
Granger C.V., Fiedler R.C., Wright, B.D. The Placid versus Distress Measure: Outpatient Physiatric Follow-Along Part III. **American Journal of Physical Medicine and Rehabilitation.**
Thomee R., Grimby G., Wright B.D. Rasch Analysis of Visual Analog Scale Measurements Before and After Treatment of Patients with Patellofemoral Pain Syndrome. **Scandinavian Journal of Rehabilitation Medicine.**
Nordenskiold U., Grimby G., Hedberg M., Wright B.D., Linacre J.M. The Structure of an Instrument for Assessment of the Effect of Assistive Devices and Altered Work Methods in Women with Rheumatoid Arthritis.
Halper A.S., Cherney L.R., Heinemann A., Semik, P., Wright B.D. Test for Right Hemisphere Communication Problems: Evaluating Psychometric Properties. **American Speech-Language-Hearing Association.**
Prieto L., Lamarca R., Santed R., Wright B.D., Alonso J. Classical Test Theory Versus Rasch Analysis for Health-Related Quality of Life Questionnaire.
Chang C. Wright B.D. Detecting Unexpected Variables in MMPI-2 Social Introversion.

In Review

"Re-examination of Physical and Mental Health as Measured by the Rand-36/ SF-36" with Chih-Hung Chang, David Cella (Northwestern) and Ron Hays (Rand).
"Rasch Model-based Approach to the Study of Measurement Consistency of Different Language Versions of Health-Related Quality of Life Instruments" with Chih-Hung Chang and David Cella (Northwestern).
"Defining Primary Dimensions of Self-Reported Health" with Chih-Hung Chang, David Cella, Jamie Von Roenn (Northwestern) and Roland Skeel (Medical College of Ohio).
"Computerized Quality of Life Assessment for Low Literacy Patients" with Elizabeth Hahn et al (Northwestern, Cook County Hospital and University of Arizona)
"The SF-36 as an Outcome Measure for treatment trials of MS: with Jeremy Hobart and Alan Thompson (London Neurological Institute).
"Is it Possible to Assess Pre/Post Change using Different Instruments at Pre and Post?" with Ken Conrad (UIC).
"Development of a Diagnostic Motor Scale for Infants" with Suzann Campbell and Mike Linacre.
"Validity of the Test of Infant Motor Performance for Prediction of 6-, 9, and 12-Month Scores on the Alberta Infant Motor Scale" with Suzann Campbell and Thubi Kolobe (UIC) and Mike Linacre.

Appendix F: Annotated Bibliography of Wright's Key Measurement Works

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85–101 [http://www.rasch.org/memo1.htm]). Princeton, New Jersey: Educational Testing Service.

This paper, invited by Benjamin Bloom after he happened to sit next to Rasch on a flight from Stockholm to Copenhagen in 1965 (Andrich, 1995), was Ben's first presentation on measurement to a national audience. At the time, Wright thought this would likely be the end of what he had to say about Rasch's work, as Choppin and Panchapakesan were finishing their degrees and moving on.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*(1), 23–48.

The procedure named in the title was an unconditional maximum likelihood estimation method referred to as UCON and now better known as Joint Maximum Likelihood Estimation (JMLE). This method is robust in the face of large amounts of missing data, opening the door to the item banking and adaptive administration methods so commonly used in testing today. The application of this method in the estimation of Rasch model parameters was an innovation introduced by Wright that Rasch did not approve of. Potential bias in UCON estimates could usually be removed by the factor (L-1)/L, except in especially short tests. The paper, however, failed to mention the use of this factor in the associated computer program, meaning that Wright had to answer questions for years over why it was there and not reported in the journal article (Andrich, 1995). This paper is the first to present a standardized Z Rasch model fit statistic.

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].

Cited 655 times since publication (Google Scholar, 12 July 2017), and 34 times since the beginning of 2016, this paper covers the main themes of model formula-

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2

tion, estimation, fit, uncertainty, interpretation, and philosophical justification. Generalized measurement and the communication of shared meaning in a common frame of reference is emphasized in extended explications of item banking, test equating and linking, and the deployment of test networks. Under the heading of item banking's advantages, Wright addresses flexible integrations of national and local tests; criterion and norm referencing; defining variables in substantive terms; developmental coherence over time; continuous quality control of item properties; item bias; construct-level, rather than item-level, interpretations of measures; the diagnosis of individual student response patterns; the value of individual uncertainty (error) estimates for establishing precision; best test design; and tailored (adaptive) testing. Ongoing developments in the psychometrics of formative assessment and coherent alignments of classroom and high stakes accountability tests (Wilson, 2004; Gorin & Mislevy, 2013; National Research Council, 2006) are in many ways still putting into practice the ideas presented by Wright in this article.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago, Illinois: MESA Press. [Spanish translation, Wright, B. D., & Stone, M. H. (1998). *Diseño de mejores pruebas* (R. Vidal, Trans.). Mexico City, Mexico: CENEVAL.]

This book provides a start-to-finish introduction to measurement using a simple example (the Knox Cube Test) to illustrate the meaning and use of logit estimates, standard errors, and model fit statistics. Moreover, a strong theory of the construct measured enables prediction of the item difficulties and provides a practical approach to improving the instrument. This classic has 2,725 citations as of 12 July 2017, according to Google Scholar.

Wright, B. D. (1980). Foreword, Afterword. In *Probabilistic models for some intelligence and attainment tests, by Georg Rasch* [Reprint; original work published in 1960 by the Danish Institute for Educational Research] (pp. ix-xix, 185–199. http://www.rasch.org/memo63.htm). Chicago, Illinois: University of Chicago Press.

Wright contextualizes Rasch's contributions, relates them to previous and contemporaneous work by Thurstone, Luce and Tukey, and others, and provides extensive quotes from Andrich's (1997) interview with Rasch, conducted in 1979.

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

This book extends Wright and Stone's (1979) examination of dichotomous data into polytomous responses. It provides the same clarity of exposition, introduces a linear reformulation of Andrich's (1982) separation index, and includes multiple worked examples. The book has 3,458 citations as of 12 July 2017, according to Google Scholar.

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].

This paper was invited by the editors of the journal after a previous paper had criticized the use of ordinal scales in functional assessment, but had not offered a superior alternative and made no reference, critical or otherwise, to a long history of

other available methods offering the desired properties. This paper has 631 citations as of 27 May 2017 according to Google Scholar.

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

This chapter details the history of the IOMW series of meetings, notable for cutting-edge presentations of psychometrics and for the multiple software training sessions held in conjunction with the research focus. This first volume (Wilson 1992) in the *Objective Measurement* book series was followed by another four volumes (Wilson 1994; Engelhard & Wilson 1996; Wilson, Engelhard, & Draney, 1997; Wilson & Engelhard, 2000), with publication ceasing in 2000. The series has recently been revived under the title, *Advances in Rasch Measurement*: Garner, Engelhard, Fisher, & Wilson (2010); Brown, Duckor, Draney, & Wilson (2011).

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*(1), 3–24.

This article explains why Wright's early work in the 1950s and 1960s using factor analysis was so dissatisfying to him, and how Rasch's models for measurement provide a means for improving the quality and meaningfulness of quantitative research in psychology and the social sciences.

Wright, B. D. (1996). Composition analysis: Teams, packs, chains. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice, Vol. 3* (pp. 241–264 [http://www.rasch.org/memo67.htm]). Norwood, New Jersey: Ablex.

This article formulates models for organizing people as teams, packs, or chains in the face of three different kinds of challenges. Teams are best when the group agrees on the goal and is more able than the task is difficult (sports, work). Packs are best when the group members disagree and the task is very difficult (Manhattan Project). Chains are best when the task is dangerous and consensus is mandatory (bucket brigade, security).

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45, 52.

This article points out that stable units of measurement for length, area, volume, and weight are rooted in commerce and politics, and that they greatly predate the emergence of mathematics and science of measurement. Wright also recognizes that measures of temperature and pressure are indebted to the steam engine, and were not initially conceived in mathematical terms. Historical demands for fair measures are noted in 7th century Islam and in the Magna Carta. The Table 1 Anatomy of Inference ought to be required reading for anyone interested in science. The explanations of the differences between IRT and measurement theory are clear and compelling.

The first two paragraphs of the article are:

> After language, our greatest invention is numbers. Numbers make measures and maps and so enable us to figure out where we are, what we have, and how much it's worth. Science is impossible without an evolving network of stable measures. The history of measurement, however, does not begin in mathematics, or even in science, but in trade and construction. Long before

science emerged as a profession, the commercial, architectural, political, and even moral necessities for abstract, exchangeable units of unchanging value were well recognized.

Let us begin by recalling two dramatic turning points in political history that remind us of the antiquity and moral force of our need for stable measures. Next, we review the psychometric and mathematical histories of measurement, show how the obstacles to inference shape our measurement practice, and summarize Georg Rasch's contributions to fundamental measurement. Finally, we review some mistakes that the history of measurement has taught us to stop making.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65–104). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Key quote, p. 76:

It is essential to "reach beyond the data in hand to what these data might imply about future data, still unmet, but urgent to foresee. The first problem is how to predict values for these future data, which, by the meaning of inference, are necessarily missing. This meaning of missing must include not only the future data to be inferred but also all possible past data that were lost or never collected."

Abstract

A new measurement in psychology has emerged from a confluence of scientific and social forces which are producing a revolution in social science methodology. We begin by reviewing how the semiotics of C. S. Peirce revise and enrich our interpretation of S.S. Stevens' four "kinds of measurement" into a creative dynamic for the evolution of one kind of useful measurement. Then we recall two turning points in social history which remind us of the antiquity and moral force of our need for stable measures. Next we review the psychometric and mathematical histories of measurement, show how the obstacles to inference shape our measurement practice and summarize Georg Rasch's contributions. This brings us to some applications of the "new" measurement models produced by Rasch's work. Finally we review some mistakes that the history of measurement can teach us to stop making.

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

Key quote, p. 45:

The study of 'fit', particularly the identification of outstanding misfit, is our chief source of new information about the world of possible experience, our chief opportunity for discovery. The observation model by which we define what to count and the measurement model by which we construct estimates of ideal magnitudes from the crude concrete counting are the inventions of measurement. The misfits that then appear are the discoveries of measuring.

The growth of science, indeed of mind, arises out of an evolving dialogue between invention and discovery—between the reassurance that we know what we are doing because our inventions work and the provocation that we must not know everything about what we are looking for because we are surprised by what we find. Constructing variables engenders an interaction of experience and idea, a dialogue between invention and discovery, that is the life force of science and mind.

References

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

Andrich, D. (1982). An index of person separation in Latent Trait Theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives, 9*(1), 95–104. Retrieved from http://www.rasch.org/erp7.htm.

Andrich, D. (1997). Georg Rasch in his own words [excerpt from a 1979 interview]. *Rasch Measurement Transactions, 11*(1), 542–543. Retrieved from http://www.rasch.org/rmt/rmt111.htm#Georg.

Brown, N., Duckor, B., Draney, K., & Wilson, M. (Eds.). (2011). *Advances in Rasch measurement* (Vol. 2). Maple Grove: JAM Press.

Engelhard, G., Jr., & Wilson, M. (1996). *Objective measurement: Theory into practice* (Vol. 3). Norwood: Ablex.

Garner, M., Engelhard, G., Jr., Fisher, W. P., Jr., & Wilson, M. (Eds.). (2010). *Advances in Rasch measurement* (Vol. 1). Maple Grove: JAM Press.

Gorin, J. S., & Mislevy, R. J. (2013). Inherent measurement challenges in the next generation science standards for both formative and summative assessment (K-12 Center at Educational Testing Service No. Invitational Research Symposium on Science Assessment). Princeton: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/gorin-mislevy.pdf.

National Research Council. (2006). In M. R. Wilson & M. W. Bertenthal (Eds.), *Systems for state science assessment*. Washington, DC: The National Academies Press.

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

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

Wilson, M. (Ed.). (2004). *National Society for the Study of Education Yearbooks. Part II: Towards coherence between classroom assessment and accountability* (Vol. 103). Chicago, University of Chicago Press.

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

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

Appendix G: Glossary

Adaptive Tests Assessments are said to be adaptive when the difficulties of the questions asked are adjusted to match the abilities of the persons measured. Paper and pencil adaptive tests date back to the beginnings of testing. Computerized adaptive tests (CAT) make available a wide range of additional powerful tools, such as standards-referenced stopping rules. Ben Wright's early work with Bruce Choppin on item banking set the stage for later work in CAT.

Bloom, B. Bloom (1913–1999) was an educational psychologist on the faculty of the University of Chicago's Department of Education known for his contributions to a taxonomy of educational objectives, a theory of mastery learning, and the definition of the two-sigma problem. Bloom happened to be seated next to Georg Rasch on a flight from Copenhagen in the mid-1960s, which led to Bloom inviting Ben Wright to speak about measurement at the 1967 ETS conference on testing.

CAT Computer adaptive testing; see adaptive tests.

Dichotomous Item responses or ratings scored in two categories (correct/incorrect, agree/disagree, etc.).

Differential Item Functioning (DIF) DIF occurs when an item's difficulty varies for equal-ability individuals from specific, identifiable groups of people who differ by gender, ethnicity, age, or some other characteristic that ostensibly should not give them any advantage or disadvantage relative to other groups.

Error (see Uncertainty) Traditionally, the difference between the observed estimate and an unknown true value, but more recently, the range within which an item difficulty or person ability estimate lies. All measures and calibrations are estimated to within a given range of error that is defined in terms of test length or sample size. The latest edition of the International Vocabulary of Metrology (VIM) and the Guide to Uncertainty in Measurement (GUM) contrast the term "error" with "uncertainty."

Estimation The process of evaluating item properties from data to determine the location and uncertainty of person measures and item calibrations is referred

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2

to as estimation. Estimation is distinct from calculation. The latter has a single correct output, whereas the uncertainty of the former will vary depending on the information available. Estimation is accomplished by a wide variety of algorithms that vary in their complexity, ease of use, and rigor.

Fit Statistics The investigation of "fit" is the evaluation of how well the data match a model's expectations, or vice versa. Descriptive multivariate models are fit to data, with the aim of accounting for as much variation in the data as possible, as indicated via statistical hypothesis tests and significance levels. For measurement models, in contrast, the "fit" of the data to the model is evaluated in terms of "fit statistics" which indicate how well model-features (such as item steepness, etc.) are evidenced in the data. As a physicist, Wright was highly sensitive to the fact that measurement does not uncritically accept just whatever data happen to come through the door.

Guttman, L. Guttman (1916–1987) was a measurement innovator who developed a class of deterministic models requiring the same kind of relation to a construct later formulated probabilistically by Rasch. Known for the scalogram and the coefficient of reproducibility.

Instrument A tool for focusing observations on a construct to be measured. In science, instruments are typically calibrated relative to a standard unit. In the absence of such units, high quality interval comparisons may be estimated from ordinal observations.

Interval A unit of measurement characterized as maintaining an invariant quantity definition. One of the four levels of measurement famously identified by S. S. Stevens, along with nominal, ordinal, and ratio. Wright, thinking through the work of C. S. Peirce, held that the nominal, ordinal, interval, and ratio seemed to be less distinctly levels than successive parts of a continuum.

Invariance The criterion of unidimensional stability obtained when estimates retain their order and spacing across samples and/or instruments (within an expected range of uncertainty).

Item A question or statement on a test, assessment, survey, or other instrument used to prompt responses that can be recorded as observations.

Item Response Theory (IRT) A multivariate statistical approach to describing item response data. IRT is sometimes mistakenly assumed to provide the theoretical context for measurement, but Rasch, Wright, and most of their students assert fundamental differences between IRT and measurement theory. The latter, but not the former, for instance, is grounded in mathematical theory of minimally sufficient statistics, and has a long history of producing experimental evidence of interval-level measurement.

Loevinger, J. Loevinger (1918–2008) was known for her work in developmental psychology, her definition of the attenuation paradox and a coefficient of test homogeneity, and early recognition of the value of Rasch's work.

Logit A log-odds unit of measurement, so called because it is the natural logarithm of the response odds, where the latter is usually taken in dichotomous educational testing applications to be the ratio of the percentage correct to the percentage incorrect.

Measurement A process "of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity." "Measurement implies comparison of quantities or counting of entities." "Measurement presupposes a description of the quantity commensurate with the intended use of a measurement result, a measurement procedure, and a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions." (All quotes from the International Vocabulary of Metrology.)

Metrology "The science of measurement and its application, involving the traceability of local results to a reference standard through a documented unbroken chain of calibrations" (Quote from the International Vocabulary of Metrology).

Ordinal A unit of numerical comparison characterized as maintaining a basis for invariant orderings of observations. One of the four levels of measurement famously identified by S. S. Stevens, along with nominal, interval, and ratio.

Parameter Separation Obtained when parameter estimates approximate unidimensional invariance, meaning that, within the range of uncertainty, ability estimates maintain their scale locations irrespective of which items are answered and difficulty estimates maintain their scale locations irrespective of who responds to the items.

Polytomous Observations in more than two categories, typically obtained using rating scales or partial credit scoring schemes.

Rasch, Georg Danish mathematician who studied with Ronald Fisher and Ragnar Frisch in the 1930s; who accompanied Tjalling Koopmans to the Cowles Commission for Research in Economics at the University of Chicago in 1947; who contacted L. J. Savage in 1960 about bringing his work on measurement to Chicago; and whose measurement models and philosophy captivated Ben Wright.

Reliability Classically conceived as the proportion of true variance, but often mistakenly defined as a measure of the unidimensionality or internal consistency of a test.

Savage, L. J. A leader in the development of subjective and personal probability in statistics at the University of Chicago; lived from 1917 to 1971. Savage supported Wright's critique of educational statistics in the face of faculty opposition, and later introduced Wright to Georg Rasch, with whom he had become acquainted in the 1947 when both were affiliated with the Cowles Commission for Research in Economics at the University of Chicago.

Separation Reliability An alternative conceptualization of how to approach the estimation of reliability. Introduced by David Andrich and extended by Wright.

Sufficient Statistic Initially named by Ronald Fisher in a 1922 paper, and generally taken to refer to statistics that contain all relevant information about a parameter. In the Rasch measurement context, counts of correctly answered test questions, or sums of survey or assessment ratings, are sufficient statistics for person ability.

Thurstone, L. L. An early psychometrician at the University of Chicago who made foundational contributions to measurement theory but abandoned them in the face of controversy in favor of the more acceptable innovations in factor analysis he had begun.

Uncertainty Also referred to as "error," uncertainty pertains to the range within which an estimate can be confidently located. The International Vocabulary of Metrology defines uncertainty as "a non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used." In metrology that dispersion is represented as a standard deviation, but in psychometrics measurand uncertainty is estimated as statistical confidence interval based in the number of questions asked and partial credit or rating scale score groups applied.

Unidimensionality The assumption that only one underlying (latent) variable is the systematic determinant of the outcomes of an item. More generally, unidimensionality is being invoked when one is trying to measure one property at a time.

Index

M. Wilson, W.P. Fisher, Jr. (eds.), *Psychological and Social Measurement*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-319-67304-2