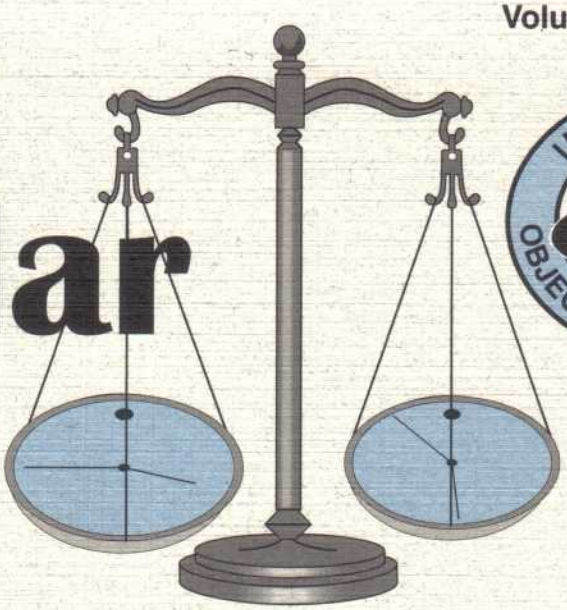
Volume 1 - No. 1- $10.00
Spring 1998

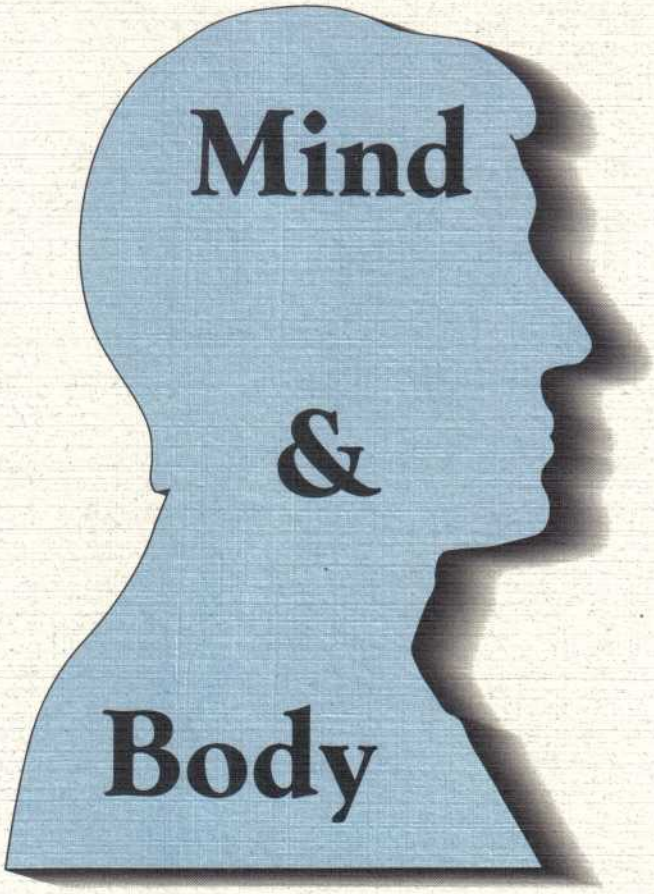
Popular Measurement



INSTITUTE FOR OBJECTIVE MEASUREMENT
1996

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Journal of the Institute for Objective Measurement



Inside:

INSTITUTE FOR OBJECTIVE MEASUREMENT
505 North Lake Shore Drive
#1308
Chicago, IL 60611

INDEX

To Whom Are We Talking?
The Need for a Primer on "Conversational" Rasch

Rita Bode
Rehabilitation Institute of Chicago

Have you ever walked in on a conversation where people were speaking another language? This happened to me when I attended an AERA session sponsored by a nonquantitative division that sounded interesting. As I sat there I realized that, although they were speaking English, I didn't have the foggiest idea what they were saying. That's what a novice must feel like when tuning into some Rasch "conversations," be they oral or written, and that is one reason why a publication such as Popular Measurement is needed. While I'm hardly an expert (I consider myself an advanced novice), I have noticed the glazed look on the faces of some audience members at Rasch presentations and thought about the need to improve our ability to communicate.

What is jargon and why do people use it? Jargon isn't just the use of specialized terminology; it also refers to the use of ordinary words that are given special meaning in certain contexts. Experts may use certain terminology to describe a complex set of phenomena or train-of-thought. When other experts use that same terminology to refer to these phenomena, jargon is created. As these descriptions become more widely known, the jargon becomes more familiar. For novices, however, the use of the jargon alone will not lead to understanding without reference to the original description of the phenomena.

Thus, within a group of experts, jargon is useful in making communication more efficient. But why do experts use jargon in other situations? There are probably many reason why they do so. They may become so accustomed to using the jargon that they forget that they acquired an understanding of it through some learning process. They assume that others have gone through the same process in understanding of the underlying phenomena. In this process, we typically acquire specific bits of information until we've collected a critical mass which enables us to understand the concept as a whole. Once we've assimilated this critical mass, we take mental shortcuts that skip over the intermediate steps. We forget that we progressed from step A to step B to step C, etc., in our acquisition and automatically leap from step A to step Z. While other experts can follow these leaps, it confuses novices who need to be lead step-by-step (as did the experts when they first acquired their knowledge) to understand new concepts.

Another reason jargon is used might be that it masks a lack of true understanding of some of the concepts involved. In the process of acquiring knowledge, certain connections may not have been made which resulted in these gaps in knowledge. If the concepts involved are truly understood, they can be explained in other terms; however, where there are gaps in understanding, one may resort to the use of jargon.

Whatever the reason for using jargon, we need to do a better job in communicating what Rasch is all about to those who don't already know about it. If conversing with Rasch experts, we can still use jargon to expand our collective understanding of new applications, but if we want to converse with novices, we need to develop bilingual skills. Conversing with novices requires the use of language which novices can understand, and contexts and examples that are relevant to them. Since there is no readily available "Rasch-to-English" dictionary, we need to develop one based on what would make sense to novices, not other experts. With the multiplicity of contexts in which Rasch is used—in education, medical rehabilitation and health sciences in general, business, etc.—multiple versions would be needed. We need to pool our resources and over time compile a list of ways of describing objective measurement to introduce new audiences.

Rita Karwacki Bode, Ph.D., has a long involvement with the development of academic achievement tests using traditional measurement theory and moving on to the development of outcome measures using Rasch measurement. She is a post-doctoral research fellow at the Rehabilitation Institute of Chicago after completion of a doctorate in Educational Psychology from the University of Illinois at Chicago.

Research Problems– Rasch Solutions

Donna Surges Tatum, Ph.D.

DECISION MAKING

We conduct research because we have questions about how to react to a given situation. The time, energy and money invested in the research and the effects of decisions require confidence in the research process. Unfortunately the complete information contained in the data does not always see the light of day. This is because traditional data analysis techniques do not access the subtleties and complexities inherent in most research situations.

We know that there are problems we should deal with when analyzing data. But because we do not know how to do so, we do the best we can with what we are used to. Today techniques enable us to address these problems directly and efficiently, instead of having nightmares about them.

RATING SCALES

RAW RATING SCALES DO NOT HAVE A UNIFORM, LINEAR STRUCTURE

Rating scales are one of the most commonly used research tools. Surveys, evaluation instruments, and psychological tests depend on ratings. Standard analyses treat these ratings as if the choices were evenly spaced steps equally separated. This is not the case.

Research shows that the spacing around rating choices are not equal. Many raters have a tendency to group their choices around the middle of the scale values. The end categories are further from the points next to them than the other categories are from each other, because some raters do not like to make extreme judgments.

Instead of the intention that each category on the scale be evenly spaced:

1 2 3 4 5 6

Reality is messier:

1		2	3	4	5	6
terrible	poor		fair	good	very good	excellent

RAW SCORES ARE NOT SUITABLE FOR ADDING AND AVERAGING

ITEMS

ALL ITEMS ARE NOT EQUAL

When surveying for such things as attitudes, speech confidence, or speaking ability, the items used are not all at the same point on the scale. Some items demand a more intense attitude than others, or a greater level of ability.

It is easier for students to agree that they are more comfortable preparing a speech than that they enjoy giving speeches. It is easier for them to demonstrate knowledge of their topic than to have good gestures.

Indeed, it would not be useful if all items did measure at the same point on the scale. That would not allow us to discover the structure of the variable. Important information is contained in the differences between elements, the difference between hard and easy items. Understanding the hierarchal structure of the items improves information for decision-making.

ITEMS MUST BE PROVEN VALID AND RELIABLE

Items must also be examined to determine whether they all relate to the same variable, or whether there are different subscales. The items must behave in a predictable manner. When some items are misunderstood by those that use the rating form, we must discover this. We must find out whether our items fit the theoretical construct we intend — the idea which motivates our research.

EXAMINE ITEMS FOR ORDER OF DIFFICULTY AS WELL AS VALIDITY

Donna Surges Tatum earned her B.A. and M.A. from Purdue University in Communication with an emphasis on Persuasion and Organizational Communication. She moved to Chicago upon graduation to join the "real" world. For seven years she worked in advertising and marketing until she realized that she was only in an alternate reality. She became a consultant and returned to academia, teaching at Roosevelt University. She was Director of the Communication Studies Department from 1986 to 1989.

Donna received her Ph.D. in 1991 from MESA at the University of Chicago. She has been teaching since 1990 in the Graham School of General Studies at the University of Chicago. It must mean something (she's not sure what) that she teaches the two courses most hated by most people: Public Speaking and Statistics.

In 1991 Donna started Meaningful Measurement, a consulting consortium for communication training, organizational development, market research, and educational assessment. Her leisure activities are swimming, yoga, and reading mystery novels.

e-mail: surgstatum@aol.com

RATERS

ALL RATERS ARE NOT EQUAL — THEY ARE INDIVIDUAL IN THE WAY THEY JUDGE A SITUATION

Raters are a crucial element in many research projects. We know from Communication and Psychology theory that we each live in our own perceptual world, and attend to our own things. One person will react more to how a speech is organized than how it is delivered. Another may be the opposite.

No matter how hard we try to train raters, we will never achieve the ideal in which all raters are the same. Instead of a false assumption of sameness, we must address the issue of differences. In fact, the real differences between raters is important additional information.

But different raters have different levels of severity when judging an event, thus we cannot take their raw scores and add them to come up with an objective measure. One rater's "3" may be worth more than another rater's "4" because that first rater is consistently more critical in her judgments. Once again we see that we cannot use the raw scores for mathematical functions.

RATERS MUST BE CONSISTENT IN THEIR JUDGMENTS

We hope that our raters are well-trained and well-behaved. But if a rater is inconsistent in judgment, then we must be able to detect who is or is not providing consistent evaluations. Otherwise we will have no basis upon which to make comparisons.

RESULTS

AN AVERAGE OR PERCENTAGE IS NOT A MEASURE

When results are given in terms of raw scores with averages or percentages, they are descriptive of one-time events. The results are not true measures because they can not be used to perform arithmetic functions such as addition, subtraction, and multiplication.

One of the fundamental errors made in research is to use scores to perform a function for which they are not equipped — to measure instead of describe. This is like using a "rubber ruler;" there is no consistency or comparability between persons, items, or groups. Scores describe a one-time event, after which the rubber ruler has to be thrown away because it is of no further use. It is not a calibrated ruler of units with fixed intervals. There is no common frame of reference with standardized measures. Subsequent research will be "measured" with another rubber ruler that is not really the same thing, even though the appearance is the same. This leads to fuzzy descriptions instead of facts of measurement.

DIRECT COMPARISONS REQUIRE A STRAIGHT LINE

Without a straight line marked in equal intervals, direct comparisons lack precision and accuracy. Tracking products over time, from group to group, or in field tests can be tedious, difficult, and imprecise. If a calibrated ruler is used to measure instead of a rubber ruler, then pictures and maps can be drawn to show the results. A well-drawn picture is worth a thousand numbers. It creates perspective.

A STABLE FRAME OF REFERENCE MUST BE CREATED AND MAINTAINED TO MAKE MEANING OUT OF DATA

SOLUTION

Many years of careful research produced a scientific method based on the Rasch Model. This system for research and data analysis is Objective Measurement. In 1953 Georg Rasch, a Danish mathematician, was hired by the Danish government to develop achievement tests to place army recruits. He discovered a mathematical model that was completely different from any used previously for this type of data analysis. In 1960 Rasch came to the University of Chicago for a year where he met Benjamin D. Wright. Professor Wright, a psychologist who originally trained as a physicist, saw the implications of this method. In 1963 he founded the MESA Psychometric Laboratory at the University of Chicago where he and his colleagues refined and extended the Rasch model. In the process they revolutionized social science research.

METHOD IN BRIEF

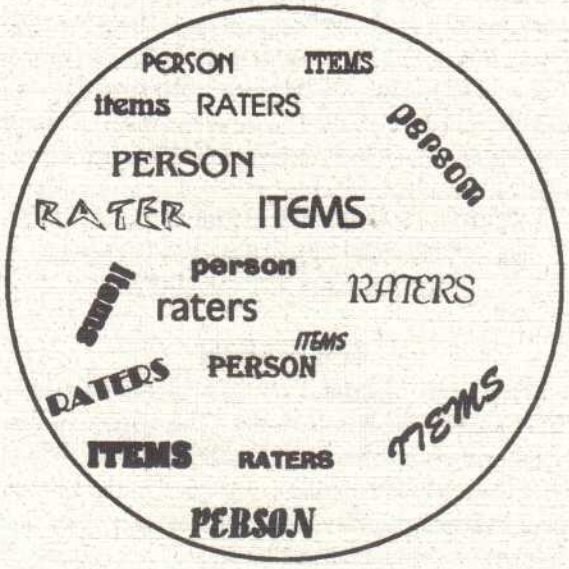
This is a brief explanation of the concepts inherent to understanding Objective Measurement. This unique approach to rater-mediated evaluations provides the most objective means for assessment yet discovered.

The Research Situation:

A traditional analysis of raw scores is primarily descriptive. It gives us a simple snapshot of the research situation. It portrays a specific group of people using a particular set of test items at a given time. All the elements are inextricably bound together. Raw scores are not linear, and do not have the mathematical properties of true measurement.

Social scientists take a snapshot of the research situation as represented by the circle below. They or others replicate the snapshot and then compare snapshots. However, these circles are not directly comparable. Each one is unique unto itself. Each circle reflects a particular, discrete situation. Averages, percentages, or percentiles based on raw scores are sample dependent, and can only represent what is happening

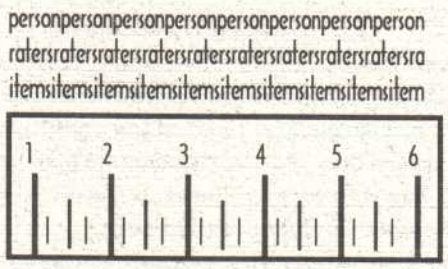
in **that** circle with **those** elements at **that** time. The results are not a measure that transcends from the particular to the general.



Measured Elements

When raw scores are conditioned using Objective Measurement techniques, something wondrously useful occurs. The strands in the analysis are disentangled from each other, and smoothed out into straight lines. They are calibrated into common units, providing context-free rulers that are able to measure at any time and any place. These results are precise reproducible measurement instead of fuzzy idiosyncratic descriptions of statistics.

Investigation is now possible in a manner that conforms to scientific principles. Instruments are constructed and calibrated to produce generalizable results. Each element can be examined separately, allowing us to delve into the data in a far deeper way than has been possible with traditional methods. We discover information heretofore unavailable.



This is it in a nutshell:

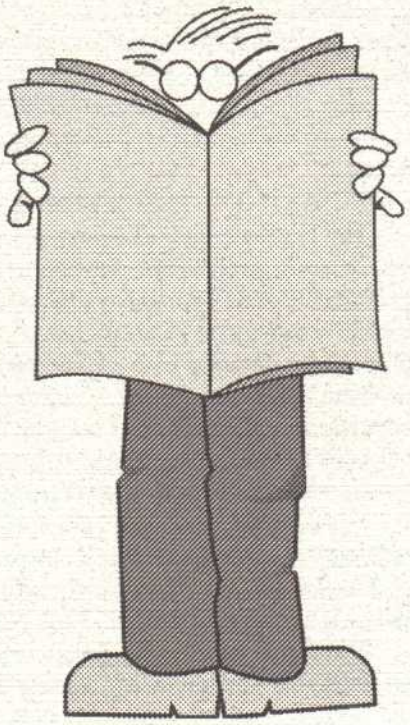
Observational statistics like raw scores and ratings describe a one-time event with all elements interwoven. Objective Measurement gives us straight lines, precise measures, and separated elements that remain stable across time and sample.

THE LEXILE FRAMEWORK FOR READING

This Lexile Framework for Reading helps you to match your Lexile measure to literature titles and everyday world texts such as USA Today. Your reading measure is determined by locating the text measure in Lexiles you can read with 75% comprehension. In other words, if you can read The Old Man and the Sea measured at 900 Lexiles, and answer correctly 75 out of 100 questions about it, you can read at 900L. Each entry on this map has been measured to determine its location.

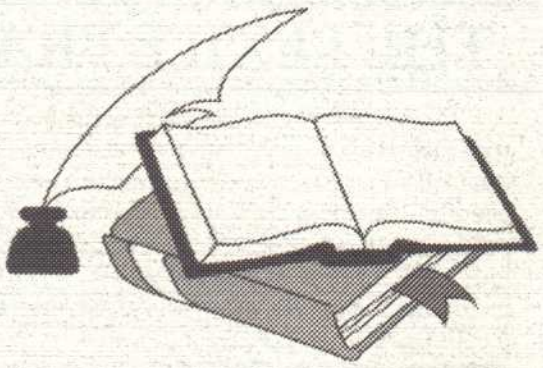
200	Ronald Morgan Goes to Bat
260	One Fish, Two Fish, Red Fish, Blue Fish
300	Mog - The Forgetful Cat
350	Little Rabbit
380	Tales of a Fourth Grade Nothing
430	Yonder
480	Curious George
530	There's a Boy in the Girls' Bathroom
560	Madeline's Rescue
620	Jack and Jill
640	The Hardy Boys: The Submarine Caper
690	How to Eat Fried Worms
730	Harriet the Spy
780	The Boy Scout Manual
810	Johnny Appleseed
830	Sounder
880	The Red Pony
920	To Kill a Mockingbird
960	The Adventures of Tom Sawyer
990	Jonathan Livingston Seagull
1040	The Pearl
1060	Dr. Zhivago
1080	USA Today
1100	Treasure Island
1120	National Geographic
1160	Trivial Pursuit Game Instruction
1200	Gulliver's Travels
1220	The Call of the Wild
1240	1040 Tax Instructions
1300	U.S. News and World Report
1340	A Brief History of Time
1360	The Odyssey
1400	The Wall Street Journal
1450	The Complete Works of Homer
1480	The Gettysburg Address
1540	The U.S. Constitution
1570	The Declaration of Independence
1630	The New England Journal of Medicine
1670	The Age of Empire
1690	Antiseptic Principles of the Practice of Surgery

First Grade
Second Grade
Third Grade
Fourth Grade
Fifth Grade
Sixth Grade
Seventh Grade
Eighth Grade
Ninth Grade
Tenth Grade
Eleventh Grade
Twelfth Grade
College Freshman-Sophmore
College Junior-Senior
Graduate School



READING RULER

A Map To Higher Levels Of Achievement



A. Jackson Stenner, Ph.D.

Student testing is a sensitive topic, one that often generates more heat than light among educators, parents, community groups, and other interested parties. By measuring students' skill levels, teachers and administrators hope to gain information that can help them to improve student performance. Unfortunately, current testing methods interpret results in terms of how the test-taker compares with other students, rather than assessing achievement against meaningful standards. Students, along with their parents and teachers, are left with the knowledge that "Johnny is at the eightieth percentile of comparison group," instead of understanding that "Johnny has achieved a desirable goal, such as being able to read USA Today."

As a result, teachers lack an objective assessment of what their students can read, and parents have only a frustratingly vague sense of whether or not their children are progressing satisfactorily.

To combat this problem, several researchers under the auspices of the National Institutes of Health have developed a unique tool that provides a clear measure of a student's reading assessment. Called the Lexile Framework, this tool assesses students according to an absolute, invariant standard, rather than merely comparing their reading performance to that of their peers. Teachers and parents receive the information they need to help students take the necessary steps to improve their reading.

WHAT IS THE LEXILE FRAMEWORK?

The Lexile Framework is an assessment system that enables educators to determine precisely a student's level of reading comprehension. The system is based on research conducted over a 15-year period by Drs. A. Jackson Stenner and Malbert Smith of MetaMetrics, Inc., Dr. Donald S. Burdick of Duke University, and faculty from the University of North Carolina, the University of Chicago, and Stanford University, with funding from the National Institutes of Health. This research, in turn, was based on more than 40 years of study by various specialists in the field of reading comprehension. In 1994, the Lexile Framework was made commercially available by MetaMetrics, Inc., an educational research and development firm based in Research Triangle Park, North Carolina.

The Lexile Framework applies well-established analytic methods to the definition of "reading comprehension." At the heart of this system is the Lexile Analyzer, a Windows-based software program that can evaluate the reading challenge of any text — books, articles, test items — by analyzing its syntactic complexity and semantic difficulty. The analyzer calibrates the text by carefully dissecting it and studying its characteristics, such as sentence length and word frequency. Unlike other readability formulas, the Lexile Framework enables you to place people and text on the same scale.

One outcome of co-calibrating text and people is a measure of reading difficulty expressed as a Lexile, a unit of measurement for reading comprehension. Longer sentence lengths and words of lower frequency lead to higher Lexile measures, since words that are unfamiliar to the reader contribute more to a text's difficulty than do familiar words. Word frequency information is derived from the five-million word corpus American Heritage Word Frequency Book by John B. Carroll, Peter Davies, and Barry Richman.

Text samples from any source — books, newspapers, standardized test items — can be calibrated simply by being scanned into a computer and imported into the Analyzer. For example, the Lexile Analyzer could be used to calibrate the contents of an entire school library. With each book's Lexile calibration included in the card catalog, librarians, teachers, and students could select materials appropriate for readers at different levels more easily and accurately.

In addition to calibrating the reading difficulty of specific text, the Framework also can be used to measure a student's reading ability. When standardized test items are calibrated, the Analyzer generates a table, called a correspondence table, that acts as a yardstick for measuring a student's level of reading comprehension. Such a correspondence table can be generated for any test, thereby providing a corresponding Lexile measure to each number correct on the test. If a student's Lexile measure is already known, the table can be used to predict a student's count correct on the test.

Students' Lexile measurements can also be determined by the Lexile Test of Reading Comprehension, which uses authentic text from published sources to assess students' reading abilities. Alternately, school systems can construct their own tests using the Lexile Analyzer.

"The Lexile Framework standards are literature-based, making the Framework uniquely useful to educators and par-

ents," says Dr. William J. Brown, Jr., an assessment specialist and former director of testing with the North Carolina Department of Public Instruction. "All other reading tests require you to interpret results in terms of how the test-taker compares to others. Because the frame of reference is the normative group, the ruler by which you're measuring is made of rubber — it bends as the cohort changes."

In contrast, notes Brown, the Lexile Framework creates an absolute standard that is embedded in the ability to read the text, and measures the ability of the test-taker by his performance against those reading standards.

"You might compare it to the President's Physical Fitness Test," says Brown. "A child is expected to do so many push-ups and pull-ups or run a certain distance in a certain time, and that tells you how fit he is ad what he needs to do to increase his level of fitness. In the same way, if you know that a student is reading at 700 Lexiles, you understand what level of material he's mastered and what books you could recommend that would help him to improve his academic skills."

The production of recommended reading lists is another unique benefit of the Lexile Framework. Through a component of the system called the Lexile Report Generator, parents and teachers receive students' Lexile measures with examples of what they can read, along with student-specific lists of books whose Lexile measurements are appropriate for readers at that level. These recommended materials are an ideal match for a student's current ability — neither so easy as to bore nor so difficult as to frustrate the student. Students and their parents and teachers are presented with a clear path to improved reading comprehension.

In addition, a richly annotated Lexile Map provides an extensive list of texts, from novels and nonfiction books to newspapers and magazines, at various levels of Lexile measurement. This color-coded poster-sized graphic makes it easy to "see" how reading develops and to select other reading materials as students progress in their reading comprehension.

"The Lexile Framework manifests what good teachers try to do anyway, which is to judge where a student is and find material that will challenge him adequately without being so difficult that he loses his motivation," says Brown. "The problem is that as children get into the latter stages of elementary school, the variance in texts and among students increases dramatically. The choice of material expands and the range of reading skills widens, so it becomes much harder for teachers to make accurate judgments about where children are and what materials are good choices for them. By using the Lexile Framework, schools can take the guesswork out of this equation, and operationalize the selection of developmentally appropriate material for their students."

The benefits for families are no less important. By giving parents an accurate assessment of their children's achievements and recommending specific materials to enhance their skills, the Lexile Framework can relieve parents' frustration and confusion and make them active partners with the teacher in students' academic progress.

"Most teachers will tell you that trying to explain to parents a child's test results in percentiles is their worst nightmare," says Brown. "Saying that little Julie is in the sixty-fifth percentile for her grade is too vague for a lot of parents. It's the kind of 'eduspeak' that can confuse and possibly alienate families instead of bringing them into the educational process. What parents want to know is, 'How is my child doing? Is she learning what she needs to learn and moving forward at a steady pace? And what can I do to help her?' With the Lexile Framework, parents get firm answers to these questions and concrete suggestions for helping their child."

In addition, Lexile measurements can help students themselves to take a more active role in their own learning, by giving them a clear picture of their abilities and a map they can follow to increase their reading comprehension.

"If you can say to a student, 'You are reading at 900 Lexiles, so a good choice for you now would be Hemingway's *The Old Man and the Sea*. When you master that, you'll be ready for *Twenty Thousand Leagues Under the Sea* or *The Hobbit*,' it gives him the idea that he has an important role to play in his own progress. Having a clear-cut path to follow encourages him to move forward and succeed."

Using score-to-measure correspondence tables, the Report Generator can forecast a student's performance on standardized tests such as the Scholastic Aptitude Test (or SAT). This "advance warning" can give students the information and incentive to achieve the levels of mastery needed for optimal performance on critical tests like the SAT.

" 'Empowerment' has become a hackneyed word, but that's the key advantage of the Lexile Framework — it gives students, parents teachers and administrators accurate information that empowers them," says Brown. "With a Lexile measure, you know precisely where a student stands in terms of an absolute scale of reading comprehension, and you know exactly what steps that student needs to take to reach higher levels of performance."

A. Jackson Stenner, Ph.D.

Jack Stenner is co-founder and Chairman of MetaMetrics, Inc.

MetaMetrics is a privately held corporation that specializes in research and development in the field of education. He has been Principal Investigator on five grants from the National Institute of Health, (1984-1996) dealing with the measurement of literacy.

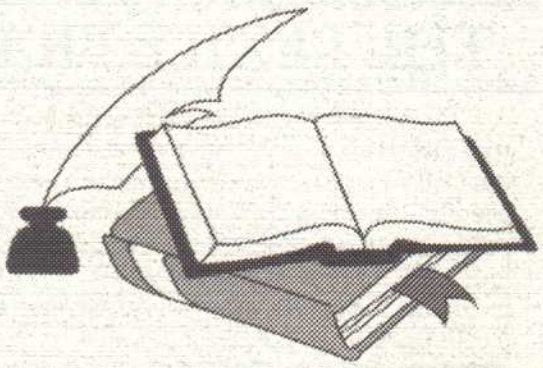
Jack Stenner is also former Chairman and co-founder of National Technology Group, a 700-person firm specializing in computer networking and systems integration which was sold to VanStar Corporation in December 1996.

He holds a Ph.D degree from Duke University and Bachelor degrees in Psychology and Education from the University of Missouri.

Jack is President of the Institute for Objective Measurement in Chicago, Illinois. He serves as a board member for The National Institute for Statistical Sciences (NISS) and is Immediate Past President of the Professional Billiard Tour Association (PBTA).

Jack resides in Chapel Hill, North Carolina with his wife, Jennifer, and their four sons.

A Map To Higher Levels Of Achievement



A. Jackson Stenner, Ph.D.

Student testing is a sensitive topic, one that often generates more heat than light among educators, parents, community groups, and other interested parties. By measuring students' skill levels, teachers and administrators hope to gain information that can help them to improve student performance. Unfortunately, current testing methods interpret results in terms of how the test-taker compares with other students, rather than assessing achievement against meaningful standards. Students, along with their parents and teachers, are left with the knowledge that "Johnny is at the eightieth percentile of comparison group," instead of understanding that "Johnny has achieved a desirable goal, such as being able to read USA Today."

As a result, teachers lack an objective assessment of what their students can read, and parents have only a frustratingly vague sense of whether or not their children are progressing satisfactorily.

To combat this problem, several researchers under the auspices of the National Institutes of Health have developed a unique tool that provides a clear measure of a student's reading assessment. Called the Lexile Framework, this tool assesses students according to an absolute, invariant standard, rather than merely comparing their reading performance to that of their peers. Teachers and parents receive the information they need to help students take the necessary steps to improve their reading.

WHAT IS THE LEXILE FRAMEWORK?

The Lexile Framework is an assessment system that enables educators to determine precisely a student's level of reading comprehension. The system is based on research conducted over a 15-year period by Drs. A. Jackson Stenner and Malbert Smith of MetaMetrics, Inc., Dr. Donald S. Burdick of Duke University, and faculty from the University of North Carolina, the University of Chicago, and Stanford University, with funding from the National Institutes of Health. This research, in turn, was based on more than 40 years of study by various specialists in the field of reading comprehension. In 1994, the Lexile Framework was made commercially available by MetaMetrics, Inc., an educational research and development firm based in Research Triangle Park, North Carolina.

The Lexile Framework applies well-established analytic methods to the definition of "reading comprehension." At the heart of this system is the Lexile Analyzer, a Windows-based software program that can evaluate the reading challenge of any text — books, articles, test items — by analyzing its syntactic complexity and semantic difficulty. The analyzer calibrates the text by carefully dissecting it and studying its characteristics, such as sentence length and word frequency. Unlike other readability formulas, the Lexile Framework enables you to place people and text on the same scale.

One outcome of co-calibrating text and people is a measure of reading difficulty expressed as a Lexile, a unit of measurement for reading comprehension. Longer sentence lengths and words of lower frequency lead to higher Lexile measures, since words that are unfamiliar to the reader contribute more to a text's difficulty than do familiar words. Word frequency information is derived from the five-million word corpus American Heritage Word Frequency Book by John B. Carroll, Peter Davies, and Barry Richman.

Text samples from any source — books, newspapers, standardized test items — can be calibrated simply by being scanned into a computer and imported into the Analyzer. For example, the Lexile Analyzer could be used to calibrate the contents of an entire school library. With each book's Lexile calibration included in the card catalog, librarians, teachers, and students could select materials appropriate for readers at different levels more easily and accurately.

In addition to calibrating the reading difficulty of specific text, the Framework also can be used to measure a student's reading ability. When standardized test items are calibrated, the Analyzer generates a table, called a correspondence table, that acts as a yardstick for measuring a student's level of reading comprehension. Such a correspondence table can be generated for any test, thereby providing a corresponding Lexile measure to each number correct on the test. If a student's Lexile measure is already known, the table can be used to predict a student's count correct on the test.

Students' Lexile measurements can also be determined by the Lexile Test of Reading Comprehension, which uses authentic text from published sources to assess students' reading abilities. Alternately, school systems can construct their own tests using the Lexile Analyzer.

"The Lexile Framework standards are literature-based, making the Framework uniquely useful to educators and par-

ents," says Dr. William J. Brown, Jr., an assessment specialist and former director of testing with the North Carolina Department of Public Instruction. "All other reading tests require you to interpret results in terms of how the test-taker compares to others. Because the frame of reference is the normative group, the ruler by which you're measuring is made of rubber — it bends as the cohort changes."

In contrast, notes Brown, the Lexile Framework creates an absolute standard that is embedded in the ability to read the text, and measures the ability of the test-taker by his performance against those reading standards.

"You might compare it to the President's Physical Fitness Test," says Brown. "A child is expected to do so many push-ups and pull-ups or run a certain distance in a certain time, and that tells you how fit he is ad what he needs to do to increase his level of fitness. In the same way, if you know that a student is reading at 700 Lexiles, you understand what level of material he's mastered and what books you could recommend that would help him to improve his academic skills."

The production of recommended reading lists is another unique benefit of the Lexile Framework. Through a component of the system called the Lexile Report Generator, parents and teachers receive students' Lexile measures with examples of what they can read, along with student-specific lists of books whose Lexile measurements are appropriate for readers at that level. These recommended materials are an ideal match for a student's current ability — neither so easy as to bore nor so difficult as to frustrate the student. Students and their parents and teachers are presented with a clear path to improved reading comprehension.

In addition, a richly annotated Lexile Map provides an extensive list of texts, from novels and nonfiction books to newspapers and magazines, at various levels of Lexile measurement. This color-coded poster-sized graphic makes it easy to "see" how reading develops and to select other reading materials as students progress in their reading comprehension.

"The Lexile Framework manifests what good teachers try to do anyway, which is to judge where a student is and find material that will challenge him adequately without being so difficult that he loses his motivation," says Brown. "The problem is that as children get into the latter stages of elementary school, the variance in texts and among students increases dramatically. The choice of material expands and the range of reading skills widens, so it becomes much harder for teachers to make accurate judgments about where children are and what materials are good choices for them. By using the Lexile Framework, schools can take the guesswork out of this equation, and operationalize the selection of developmentally appropriate material for their students."

The benefits for families are no less important. By giving parents an accurate assessment of their children's achievements and recommending specific materials to enhance their skills, the Lexile Framework can relieve parents' frustration and confusion and make them active partners with the teacher in students' academic progress.

"Most teachers will tell you that trying to explain to parents a child's test results in percentiles is their worst nightmare," says Brown. "Saying that little Julie is in the sixty-fifth percentile for her grade is too vague for a lot of parents. It's the kind of 'eduspeak' that can confuse and possibly alienate families instead of bringing them into the educational process. What parents want to know is, 'How is my child doing? Is she learning what she needs to learn and moving forward at a steady pace? And what can I do to help her?' With the Lexile Framework, parents get firm answers to these questions and concrete suggestions for helping their child."

In addition, Lexile measurements can help students themselves to take a more active role in their own learning, by giving them a clear picture of their abilities and a map they can follow to increase their reading comprehension.

"If you can say to a student, 'You are reading at 900 Lexiles, so a good choice for you now would be Hemingway's *The Old Man and the Sea*. When you master that, you'll be ready for *Twenty Thousand Leagues Under the Sea* or *The Hobbit*,' it gives him the idea that he has an important role to play in his own progress. Having a clear-cut path to follow encourages him to move forward and succeed."

Using score-to-measure correspondence tables, the Report Generator can forecast a student's performance on standardized tests such as the Scholastic Aptitude Test (or SAT). This "advance warning" can give students the information and incentive to achieve the levels of mastery needed for optimal performance on critical tests like the SAT.

" 'Empowerment' has become a hackneyed word, but that's the key advantage of the Lexile Framework — it gives students, parents teachers and administrators accurate information that empowers them," says Brown. "With a Lexile measure, you know precisely where a student stands in terms of an absolute scale of reading comprehension, and you know exactly what steps that student needs to take to reach higher levels of performance."

A. Jackson Stenner, Ph.D.

Jack Stenner is co-founder and Chairman of MetaMetrics, Inc.

MetaMetrics is a privately held corporation that specializes in research and development in the field of education. He has been Principal Investigator on five grants from the National Institute of Health, (1984-1996) dealing with the measurement of literacy.

Jack Stenner is also former Chairman and co-founder of National Technology Group, a 700-person firm specializing in computer networking and systems integration which was sold to VanStar Corporation in December 1996.

He holds a Ph.D degree from Duke University and Bachelor degrees in Psychology and Education from the University of Missouri.

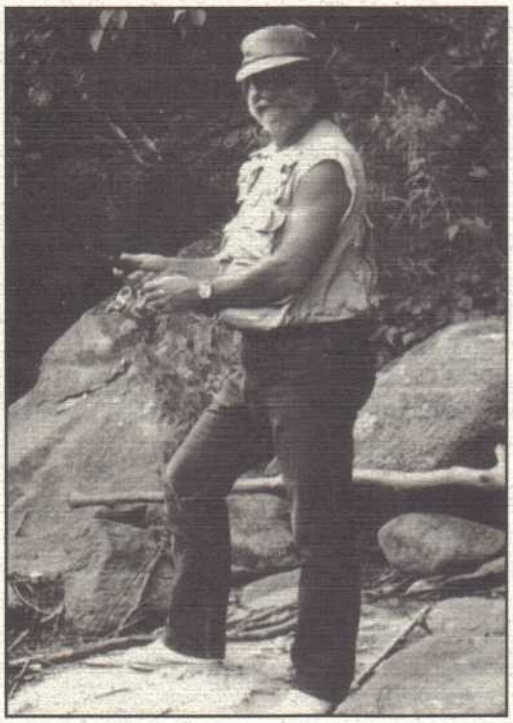
Jack is President of the Institute for Objective Measurement in Chicago, Illinois. He serves as a board member for The National Institute for Statistical Sciences (NISS) and is Immediate Past President of the Professional Billiard Tour Association (PBTA).

Jack resides in Chapel Hill, North Carolina with his wife, Jennifer, and their four sons.

GALTON: The First Psychometrician?

Larry H. Ludlow, Ph.D. - Boston College

Ever wonder how many brush strokes it takes to create a painting? Or how to measure boredom, attraction to the opposite sex, the efficacy of prayer, or the intelligence of earthworms? Sir Francis Galton wondered about these things and set out to develop procedures and instruments by which such questions could be answered and replicated. In fact, he counted *everything* that appeared to have any form of regularity.



Larry H. Ludlow

He counted brush strokes while sitting for his own portrait at two different times in his life. Karl Pearson suggested his "pained" expression was due to his concentration while counting. It took about 24,000 strokes for each painting.

He counted spikes of flowers on trees. By counting the spikes of flowers on a typical tree, and then the number of trees along a one mile stretch of road, he estimated that the number "one million" could be represented as the flowers on a row of trees ten miles in length.

He counted the fidgets of persons sitting through a boring lecture. He investigated the "instances in which men who are more or less illustrious have eminent kinfolks." This was the basis for his argument that genius is hereditary (Galton, 1869). One conclusion was that great commanders tend to be small because their relative chance of being shot varies as the square root of the product of their height and weight.

When looking at facial features, he wondered whether persons with differentiated mental characteristics also have differentiated physical features. He actually attempted the development of composite portraits for "ideal criminal" classes. He also looked at the numbers of attractive, indifferent, and repellent-looking women. The objective was to form a "Beauty Map" of the British Isles.

Galton's work produced many "firsts." His investigation of points of similarity between twins was the first use of control groups in psychological research. His research into variations in weather conditions resulted in the first published meteorological maps of Europe. His work on fingerprint characteristics led to the legal use of fingerprint identification.

He counted earthworms on a rainy sidewalk when he was helping Charles Darwin investigate the intelligence of worms. He examined the degrees of vividness of mental imagery, and the instances of phantasmagoria, causes of snoring, and on and on. He seems to have always carried a notebook and some type of ingenious device capable of pricking a piece of paper by which he recorded, unobtrusively, various aspects of events occurring around him. He even performed arithmetic by taste and smell.

What, you might reasonably ask, is the purpose of this article? It was written because it provides some relatively obscure, yet fascinating, information on the early history of psychometrics. For some years now I have taught a course in psychometrics. An important feature of the material covered in the course is the historical context within which the models and methods we employ have evolved. However, my lectures never included anything about Galton other than his development of regression and correlation. A little-appreciated fact is that Galton's original version of regression analysis consisted of reading the "inclination" off a graph of medians, labeling it *r* as a coefficient of "reversion," and then using it as an "index of co-relation." Correlation, as we know it, was actually a byproduct of regression. (See Pearson, 1930, Vol. IIIA, Chap XIV.).

My approach to the history of psychometrics is fairly standard. It begins with the classical German psychophysics of the 1800's with Weber, Wundt, and Fechner, moves into the 1900's ability testing movement with Cattell, Binet, and Spearman, and then into the psychological scaling methods associated with Thurstone. Modern test theory texts are introduced where standard presentations include something like "the field of psychometrics has a history of growth and development extending over some 75 years since the early work of Binet in France and Spearman in England" (Thorndike, 1982, p 1). And "psychometric methods" is simply defined as "procedures for psychological measurement" (Guilford, 1954, p 1). Standard stuff.

But, while working on a project tracing the role that residuals have played in the evolution of scientific models, I stumbled across some early research of Galton's. Practically everything a reasonable (or obsessed) person might want to know about Galton appears to be covered in the four volumes

of The Life, Letters and Labours of Francis Galton by Karl Pearson. In particular I became intrigued with his reference to "psychometric experiments" and I subsequently set out to track down the original use of the word "psychometrics." That effort resulted in this paper.

Galton's interests in mental operations led him to propose a "new instance of psychometry" (Galton, 1879, p 149). In his article, "Psychometric Experiments," he defined "psychometry" as the "art of imposing measurement and number upon operations of the mind." He then argued that "until the phenomena of any branch of knowledge have been subjected to measurement and numbers, it cannot assume the status or dignity of a science."

There are two interesting points in these quotes. First, I assumed psychometry was simply a term coined by Galton and that it represented some transference of Galton's experiences in the German psychophysics labs to the realm of "mind." It turns out that there was a "science of psychometry" during the mid-to-late 1800's devoted to the investigation of mental divining of qualities and properties of objects or persons by a "psychometrician" (Buchanan, 1854).

Second, his quote is remarkably similar in spirit to William Thomson's circa 1883 famous dictum about measurement and science. See Merton et al. for what Baron Kelvin of Largs, or Lord Kelvin (William Thomson at the time) said, and how and why it differs from what is engraved in the facade of the University of Chicago Social Science research building. Actually, the statement's sentiment can be traced back to John Arbuthnot (1692). His work illustrated what he called the psychometric side of anthropology.

For his 1879 article Galton repeated an experiment in "mental operations" four times, under different circumstances, at intervals of about one month. The experiment consisted of recording the "thoughts arisen through direct association" with a list of 75 words. He did not publish his lists because "they lay bare the foundations of a man's thoughts with curious distinctness, and exhibit his mental anatomy with more vividness and truth than he would probably care to publish to the world." This is a good example of the honest and open writing style so characteristic of the period. In other words, he conducted experiments in what we now call free-association. This could well be the earliest investigation of free-association, a psychoanalytic technique developed from the 'talking cure' and Freud's interpretation of dreams (Berg and Pennington, 1966, p 594)). He threw his resulting thoughts into a "common statistical hotch-pot" (This sounds like our word "hodgepodge" and our analysis called the "shotgun approach"). Galton determined (a) the rate at which ideas were formed (50 per minute), (b) the frequency of recurrent associations (about one half), (c) the frequency within periods of his life that associations could be attributed (showing "in a measurable degree, the large effect of early education in fixing our associations"), and (d) the character of associations that occurred (verbal, sensory, "histrionic").

The significance of this article is that it is, I believe, the first published investigation in the field that we presently know as psychometrics. Although he had notes titled "Psychometric Inquiries 1876," and published "Psychometric Facts" in Nineteenth Century, March 1879, p 425-33, they were not of a statistical nature. Granted, Galton's psychometric research differs somewhat from what we, as psychometricians, typically mean when we say we are conducting psychometric analyses, but his work is compatible with our current approach to psychometrics. That is, psychometrics is the quantification of psychological phenomena.

What else does Galton have to offer? When addressing mental tests he states:

"There are many faculties that may be said to be potentially constant in adults though they are not developed, owing to want of exercise. After adequate practice, a limit of efficiency would in each case be attained and this would be ***a personal constant*** *(emphasis added); but it is obviously impossible to guess what that constant would be from the results of a single trial. No test professes to do more than show the efficiency of the faculty at the time it was applied, and many tests do even less than this"* (Galton (1885), in Pearson, Vol. II, pp. 371-2).

This quote contains the kernel of the classical true-score concept, including notions of reliability and validity. Note also that the quote appears 20 years earlier than the seminal work on measurement error by Spearman. Galton, the first psychometrician?..Yes.

References:

Berg, I.A. & Pennington, L.A. (1954). (3rd ed.) *An Introduction to Clinical Psychology*. NY: Ronald.

Buchanan, R. (1854). Lectures in Neurological Systems of Anthropology. p 124

Galton, F. (1869). *Hereditary Genius*. London: Macmillan.

Galton, F. (1879). Psychometric experiments. *Brain: A Journal of Neurology, II*, 149-162.

Guilford, J.P. (1954). (2nd ed.). *Psychometric Methods*. NY: McGraw-Hill.

Merton, R.K. Sills, D.L. & Stigler, S.M. (1984). The Kelvin Dictum and social science: An excursion into the history of an idea. *Journal of the History of the Behavioral Sciences*, 20, 319-331.

Pearson, K. (1914, 1924, 1930a, 1930b). *The Life, Letters and Labours of Francis Galton*. Vol. I, II, IIIA, IIIB. Cambridge: Cambridge University Press.

Thorndike, R.L. (1982). *Applied Psychometrics*. Boston: Houghton.

I gratefully acknowledge the persistent literature search efforts expended by Susan Henderson-Conlon.

Larry H. Ludlow, Ph.D.

Associate Professor, Boston College, School of Education, Education Research, Measurement, and Evaluation Program.

Professional interests: developing interesting graphical representations of multivariate data (visualizing an eigenvector), and applying psychometric models in situations where the results have an obvious practical utility (scaling flute performance).

Personal interests: woodcarving, sketching, and motorcycling.

Last book read: Arthur Koestler, The Sleepwalkers.

Personal goal: Actually catch something fly-fishing.

Favorite drink: Diet Dr. Pepper.

Favorite quote: "If it exists, it can be measured. If it can't be measured, it doesn't exist." (mine)

e-mail: LUDLOW@BC.EDU

GEORG RASCH

The Man Behind The Model

The Mathematician

Georg Rasch

Benjamin Drake Wright, Ph.D.

Georg Rasch, Doctor of Philosophy in mathematics (1930), member of the International Statistical Institute (1941), charter member of the Biometrics Society (1947), Professor of Statistics at the University of Copenhagen (1962), and Danish Knight of the Order of Dannebrog (1967), was born in Odense, Denmark, on 21 September 1901, the youngest and "least practical" of three brothers.[1]

His mother was ill throughout his childhood and Rasch had few recollections of her. But his fiercely religious father left deep and lasting impressions. Wilhelm Rasch, sailor, ship's officer, mathematics teacher and self-anointed missionary, was, "the most hard-boiled evangelist I have ever known."

Wilhelm dragged his family to Svendborg in 1906 to open a mission high school for prospective seamen. In 1914 Georg became fascinated by the trigonometry texts in his father's library and fell in with a school teacher who made mathematics "something with which a wonderful world was opened."

The teacher realized that Georg was a born mathematician and persuaded his frugal father to invest in sending Georg to the cathedral school in Odense where there was a good mathematics curriculum. Georg made the most of it and went on to the University of Copenhagen in 1919.

> *I entered the Faculty of Science, to which mathematics belonged, and got into immediate contact with my teachers. I had, of course, to learn the elements of function theory and even geometry, but I concentrated upon the analytic part. What caught my interest was the theory of Lagrange equations. This resulted in my first publication (Neilsen & Rasch 1923).*
>
> *I got a stipend for my studies and became a member of college Regensen where we received free room and board. Since I did not see any further reason for doing arithmetical work for my living, I left Professor Neilsen and got another teacher, Professor Nørlund, who had written an extremely good book on difference equations.*
>
> *Nørlund was my professor for the rest of my time as a student, and I was his assistant teacher from 1925, when I graduated, until 1940. The topics in function theory that Nørlund lectured about together with the other topics I had to study in order to lecture as his assistant built up my mathematical background.*
>
> *Nørlund was also director of the Geodetic Institute to which I became attached to provide mathematical and computational assistance. This added to my income and in 1928, I married my sweetheart, Elna Nielsen, with the charming nickname "Nille". Two daughters were added to the family in 1931 and 1933.*
>
> *My thesis, defended in 1930, was the fruit of my cooperation with Nørlund, but in a field which he himself did not cultivate. It dealt with matrix algebra and its applications to linear systems of differential equations. I have always loved to think, but I have never*

been inclined to do much reading. So I had never seen anything about matrices. Nørlund gave lectures on difference equations in which he wrote out every equation in detail every time. When working through my notes I discovered, to my surprise, that these long equations could be condensed in a simple way. I did not know anything about matrices at that time, but just invented them for myself and discovered what their rules must be. Only later did I find out that others had formalized the same idea.

I invented my own theory of matrices, especially as they applied to linear systems of differential equations. The part of my thesis on the theory and application of product integrals which developed a linear system of differential equations as a generalization of the ordinary elementary integral was published in German (Rasch 1934). Years later I learned that the techniques developed in this paper played a part in solving problems in atomic theory and were also used to prove some difficult theorems in group theory.

The early 1930s were difficult. Aside from teaching as Nørlund's assistant and small jobs for the Geodetic Institute, there was no work in mathematics. So Rasch helped two medical acquaintances studying the reabsorption of cerebrospinal fluid to understand their data. This gave him his first experience with the exponential distribution and material for his first experimental paper (Fog, Rasch & Stürup 1934).

The success of this collaboration motivated Fog and Stürup to engage Rasch to teach mathematics and statistics to a small group of psychiatrists and neurologists. Word of this got to the head of the Hygienic Institute, who was also interested in statistics. The outcome was that Rasch served the Hygienic Institute as statistical consultant from 1934 to 1948 and also become attached to the State Serum Institute, a relationship which continued until 1956.

About the same time Nørlund, for whom Rasch still taught mathematics, and Madsen, Director of the Serum Institute, got into a conversation about Rasch's work and decided that to do his job at the Serum Institute, he needed to learn the latest developments in statistics. They applied to the Rockefeller Foundation for Rasch to study with R.A. Fisher.

The Rockefeller fellowship was granted, but, while it was brewing, Rasch went to Oslo on a Carlsberg grant to study Ragnar Frisch's confluence analysis, a technique developed for economics, but similar to factor analysis. Then in September 1934 Rasch joined Fisher at the Galton Laboratory in London.

I went through Fisher's statistical methods and soon got hold of his 1922 paper where he developed his theory of maximum likelihood. What caught my interest was his idea that this is a form of generalization of the same kind as Gauss attempted when he invented least squares.

The meaning of least squares is not, in Fisher's interpretation, however, just a minimization of a sum of squares. It is a maximization of the probability of the observations. There is an essential difference between this and the simple idea of minimizing sums of squares.

This philosophy went further when Fisher got to his concept of sufficiency. To mathematical minds sufficiency may appeal as nothing more than a surprising nice property, extremely handy when accessible, but, if not, then you just do without it. But to me sufficiency means much more than that. When a sufficient estimate exists, it extracts every bit of knowledge about a specified feature of the situation made available by the data as formalized by the chosen model. 'Sufficient' stands for 'exhaustive' as regards the feature in question.

What is left over when a sufficient estimate has been extracted from data is independent of the trait in question and may therefore be used for a control of the model that does not depend on how the actual estimates happen to reproduce the original data. This is the cornerstone of the probabilistic models that generate specific objectivity.

The realization of the concept of sufficiency, I think, is a substantial contribution to the theory of knowledge and the high mark of what Fisher did. His formalization of sufficiency nails down the conditions that a model must fulfill in order to yield an objective basis for inference.

During his year in London, Rasch also discussed the problem of relative growth with Julian Huxley. Using data on crab shell structure, Rasch discovered that it was possible to measure the growth of individual crabs as well as populations.

It meant a lot to me to realize the meaning and importance of dealing with individuals and not with demography. Later I realized that test psychologists were not dealing with the testing of individuals, but were studying how traits, such as intelligence, were distributed in populations. They were making demographic studies and not studies of individuals.

Rasch began teaching statistics to biologists in the fall of 1936. Then in 1938 the director of the University of Copenhagen Psychological Laboratory learned of Rasch's interest in statistics. The director asked Rasch to give some lectures to his psychologists. The result was a connection lasting thirty years.

Rasch began his work on psychological measurement in 1945 when he helped standardize an intelligence test for the Danish Department of Defense (Rasch 1947).

In carrying out the item analysis I became aware of the problem of defining the difficulty of an item independently of the population and the ability of an individual independently of which items he had actually solved.

A friendship with Chester Bliss formed in London in 1935

brought Rasch to the United States in 1947 to participate in the founding of the Biometrics Society (Rasch 1947a) and the postwar reorganization of the International Statistical Institute. Tjalling Koopmans, a fellow student of Ragnar Frisch's confluence analysis and Fisher's sufficient statistics, invited Rasch to spend two months with the Cowles Commission for Research in Economics at the University of Chicago, where Rasch met Jimmie Savage.

> *In 1951 I was faced with a task the solution of which added a new tool to my arsenal. The Danish Ministry of Social Affairs wanted an investigation of the development of reading ability in 125 former students of public schools in Copenhagen, who in their school years had suffered from serious reading difficulties and therefore had received supplementary education in that discipline.*
>
> *For each of these students were recorded the results of repeated oral reading tests during his school years. It would be a simple task to follow the development of a student's reading ability over a number of years if the same part of the same test were used every time, but at each testing it was necessary to choose a test which corresponded to the student's standpoint, so each student was followed up with a series of tests of increasing "degrees of difficulty."*
>
> *In a concrete formulation of this problem I imagined — in good statistical tradition — the possibility that the reading ability of a student at each stage could be characterized in a quantitative way — not through a more or less arbitrary grading scale, but by a positive real number defined as regularly as the measurement of length.*
>
> *Whether this would be possible with the tests in question could not be known in advance. It had to be tried out through a separate experiment which was carried out in January 1952. In this experiment 500 students in the 3rd – 7th school year were tested with 2 or 3 of the texts used in the earlier investigation. (Rasch 1977, 58-59)*
>
> *I chose the multiplicative Poisson for the reading tests because it seemed a good idea mathematically, if it would work. It turned out that it did and so I wanted to have some motivation for using it. In order to do so, I imitated the proof of a theorem concerning a large number of independent dichotomous events, each of which had a small probability. Under these conditions the number of events becomes Poisson distributed. I took care that my imitation ended up with the multiplicative Poisson model, that is, I made sure that there was a personal factor entering into each of the small probabilities for the dichotomous outcome and that each item would have its own parameter and then I had my new model.*
>
> *I had taken a great interest in intelligence tests while working with them in 1945. It struck me that I might analyze the test we had constructed then, and which had been taken over by the Military Psychology Group.*
>
> *The first thing I did was to analyze the Raven tests. They worked almost perfectly according to the multiplicative model for dichotomous items. That was my first example using the newly discovered model. Now I compared the results of the Raven's test and the results of my analysis of the military intelligence test. The intelligence test did not conform.*
>
> *When I showed this to the head of the military psychologists he saw the point. I had talked to him about my attempts to make sense of intelligence tests by means of the model I had discovered in connection with the multiplicative Poisson. I had also told him about the Raven's tests. Now I presented the examination of the test he actually had in current use from the Psychology Laboratory. I pointed out that it seemed to consist of different groups of items with quite different kinds of subject matter.*
>
> *His immediate reaction was to call on Borge Prien who was working for the military psychologists and to give him the order that, within the next six months, before the next testing session in November 1953, to have ready a new intelligence test consisting of four different subtests, each of these to be built in such a way that they followed the requirements that Rasch demanded.*
>
> *It was remarkable. Prien actually did that in six months. He invented tests, which, when you see them, are rather surprising. He really did invent items of the same sort, from very easy to very difficult, and spaced in a sensible way. We did do some checking in the process and omitted or modified items that did not seem to be working. It was a masterpiece. Prien had been told, 'All you have to construct is four different kinds of tests, with very different subject matters and each of them should be just as good as Georg tells us that Raven's tests are.' And so he did. That was when I really began to believe in the applicability of that elementary model.*

THE BOOK

The establishment in 1955 of the Danish Institute for Educational Research brought Rasch a wealth of problems requiring clarifications, elaborations, and extensions of the principles already laid down.

> *In 1957 I gave some lectures on the researches I had done since Prien's construction of the new intelligence tests. I told about the multiplicative Poisson and about the nice little model which sorts items out from each other. My lectures were tape-recorded, and my daughter Lotte got the task of deciphering them and writing them down. She made a proper work out of it, and what she did was taken over by the Educational*

Institute, and they had it mimeographed.

At that time the institute consisted of five different departments, each with its own head. Every Friday morning the company of them, together with the director, Erik Thomsen, and I had a meeting where we discussed current matters. Thomsen organized it so that on a number of these Fridays we went through my manuscript. That clarified many points that I had been vague about. I was forced by the young fellows there to make clear what I meant.

A preliminary Danish edition of the manuscript was carefully scrutinized by the staff members of the Institute. The Danish text was transformed into English by G. Leunbach, who has also revised later additions in English. Finally, in 1960, L.J.Savage of the University of Chicago reviewed the final manuscript critically.

The outcome of the reading test experiment was beyond expectation: a statistically satisfactory analysis on the basis of a new model which represented a genuine innovation in statistical techniques!

But the understanding of what the model entails tarried several years. Then at the 1959 anniversary of the University of Copenhagen the highly esteemed Norwegian economist Ragnar Frisch — later Nobel Prize winner — came to Copenhagen to receive an honorary doctorate. I visited him the next day, and he asked me what I had been doing in the 25 years since I stayed at his institute in Oslo for a couple of months to study his new techniques of statistical analysis. I soon concentrated on the comparison of reading speeds which I proceeded to explain.

Applying my measurement model to reading speeds states that the probability that person n in a given time reads a_{ni}, words of text i is determined by the Poisson distribution.

The Poisson distribution has the important property that the sum of the two Poisson distributed variables is also Poisson distributed with a parameter which is the sum of the two parameter values.

In a class of possible outcomes of this kind where the total number of words read, a_{n+}, has a fixed value, the probability of the outcomes a_{ni} and a_{nj} conditional on the total a_{n+}, is given by dividing the two Poisson variables.

Until now Frisch had only listened politely, but now I presented a crucial point which demands a careful inspection.

When one Poisson distribution is divided into another, factors cancel, and the resulting conditional probability does not contain the person parameter. The probability that the given number of words read, a_{n+}, is composed of a_{ni} and a_{nj} words of the two tests is therefore expressed by

$$p\{a_{ni}, a_{nj} \mid a_{n+}\} = \binom{a_{n+}}{a_{ni}} \left(\frac{E_i}{E_i+E_j}\right)^{a_{ni}} \left(\frac{E_j}{E_i+E_j}\right)^{a_{nj}}$$

which is determined by the observed numbers a_{ni} and a_{nj} and by the ratio between the difficulty parameters of the two tests E_i and E_j, while it is not influenced by which person is involved. On seeing this Frisch opened his eyes widely and exclaimed: "It (the person parameter) was eliminated, that is most interesting!" And this he repeated several times during our further conversation. To which I of course agreed every time — while I continued reporting the main results of the investigation and some of my other work.

Only some days later did I all of a sudden realize what in my exposition had caused this reaction from Ragnar Frisch. And immediately I saw the importance of finding an answer to the following question: "Which class of probability models has the property in common with the Multiplicative Poisson Model, that one set of parameters can be eliminated by means of conditional probabilities while attention is concentrated on the other set, and vice versa?"

What Frisch's astonishment had done was to point out to me that the possibility of separating two sets of parameters must be a fundamental property of a very important class of models.(Rasch 1977, 63-66)

By 1953 Rasch had used a Poisson model to analyze a family of oral reading tests and with Borge Prien had designed and built a four-test intelligence battery each test of which fit the requirements of his logistic model for item analysis. Rasch discussed his concern about sample dependent estimates in his article on simultaneous factor analysis in several populations (Rasch 1953). However, his work on item analysis remained unknown outside Denmark until 1960, when he lectured in Chicago, gave a paper at the Berkeley Symposium on Mathematical Statistics (Rasch 1961), and published *Probabilistic Models*.

PREFACE to *Probabilistic Models*

For several years statistical methods have been a favorite instrument within various branches of psychology. Warnings have, however, not always been wanting. Two instances from recent literature may serve as examples.

Skinner[1] vigorously attacks the application of statistics in psychological research, maintaining that the order to be found in human and animal behavior should be extracted from investigations into individuals, and that psychometric methods are inadequate for such purposes since they deal with groups of individuals.

As far as abnormal psychology is concerned Zubin[2] expresses a similar view in stating: "Recourse must be had to individual statistics, treating each patient as a separate universe. Unfortunately, present day statistical methods are entirely group-centered so that there is a real need for developing individual-centered statistics."

Individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments tests or items or other stimuli — within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class — "measuring the same thing" — independent of which particular individuals within a class considered were instrumental for the comparison.

This is a huge challenge, but once the problem has been formulated it does seem possible to meet it. The present work demonstrates, by way of three examples from test psychology, certain possibilities for building up models meeting these demands. And it would seem quite possible to modify and extend the methods used here to cover much larger areas, but in order to investigate how far the principles go — and what should be done outside possible limits — much research is needed. It is hoped, however, that planned continuations of the present work and contributions from others will gradually enlarge the field where fruitful models can be established. (Rasch 1960, xx-xxi)

In her 1965 review Jane Loevinger wrote,

Rasch (1960) has devised a truly new approach to psychometric problems.... He makes use of none of the classical psychometrics, but rather applies algebra anew to a probabilistic model. The probability that a person will answer an item correctly is assumed to be the product of an ability parameter pertaining only to the person and a difficulty parameter pertaining only to the item. Beyond specifying one person as the standard of ability or one item as the standard of difficulty, the ability assigned to an individual is independent of that of other members of the group and of the particular items with which he is tested; similarly for the item difficulty.... Indeed, these two properties were once suggested as criteria for absolute scaling (Loevinger, 1947); at that time proposed schemes for absolute scaling had not been shown to satisfy the criteria, nor does Guttman scaling do so. Thus, Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of non-arbitrary measures. Rasch is concerned with a different and more rigorous kind of generalization than Cronbach, Rajaratnam, and Gleser. When his model fits, the results are independent of the sample of persons and of the particular items within some broad limits. Within these limits, generality is, one might say, complete. (Loevinger 1965, 151).

In the 60's I introduced a more definite version of an old epistemological concept. I preserved the name of objectivity, but since the meaning of that word has undergone many changes since its Hellenic origin and is used in everyday speech as well as scientific discourse with many different contents, I added a restricting predicate: specific.

My professional background is mathematical and statistical, not philosophical. The concept has therefore not been carved out in a conceptual analysis, but, on the contrary, its necessity has appeared in my practical activity as a statistical consultant. (Rasch 1977, 58)

It is the two earliest and most popular members of this "very important class of models" which Rasch applies in *Probabilistic Models*. Although the book focuses on the measurement of reading accuracy, speed, and intelligence, the basic principles employed are fundamental to all scientific work.

When first suggesting the models (for measuring) I could offer no better excuse for them than their apparent suitability, which showed in their rather striking mathematical properties. In Rasch (1961) a more general point of view was indicated, according to which the models were strongly connected with what seemed to be basic demands for a much needed generalization of the concept of measurement.

In continuation of that paper my attention was drawn to other fields of knowledge, such as economics, sociology, history, linguistics, evaluation of arts, etc. where claims are arising of being taken just as seriously as Natural Sciences.

On a first sight the observational material in Humanities would seem very difficult from that in physics, chemistry and biology, not to speak of mathematics. But it might turn out that the difference is less essential than it would seem. In fact, the question is not whether the observations are of very different types, but whether Sciences could be firmly established on the basis of quite different types of observation. (Rasch 1967.)

The psychometric methods introduced in Rasch's book go far beyond measurement in education or psychology. They embody the essential principles of measurement itself, the principles on which objectivity and reproducibility, indeed all scientific knowledge, are based. (Rasch 1960, xix)

THE FRIEND

One day in November 1959 Jimmie Savage asked me whether I knew a Dane named Rasch. He had encountered Rasch at a biostatistics conference in Washington. Drawing on a 1947 association in Chicago, Rasch had pressed for a return visit. He had a new way to construct objective mental measurements. Jimmie had some money for a visiting professor. If he invites Rasch, will I guarantee students? Having no control over students, I guaranteed myself.

Georg began his lectures in March 1960. At first they are jammed — most of the statistics department, quite a few social scientists, even some students. Georg is bold, dramatic, and uncompromising. He is also enthusiastically forthright about the futility of many traditional procedures. Unfortunately the statisticians are not interested in changing their ways and the social scientists find it "too mathematical." By three weeks only one "student" remains.

Nevertheless, Georg marches in each morning, sets up his notes, grasps the lectern and delivers a lecture. Then he

scans the room, focuses on his one student, steps off the platform and squeezes into the seat beside me to answer all my questions.

But it is lunchtime. In order not to interrupt our discourse, Georg invites me to his kitchen where, while continuing our animated discussion, we mash sardines into black bread with plenty of oil and black pepper and wash them down with Danish beer.

Ben Wright and Georg Rasch in Athens, Georgia, April 1973.

Why did I stay with Georg when my students and colleagues departed? Was it my promise to Jimmie? Was it my compassion for Georg? Of course. But the clincher was a dawning realization that Georg had discovered a practical solution to the most stubborn and seemingly insurmountable obstacle to any real social science, the almost complete absence of reproducible measures.

Later, as we became comfortable, I dared to tell Georg about my disappointments with the instability of the many factor analyses I had performed. Georg told me about his 1953 article on this very problem. The danger in factor analysis is that it seldom reproduces its results. But only when it can be demonstrated to have done so can it serve as a useful scientific method.

Intrigued by my failed attempts to control semantic differential data with factor analysis, Georg insists on taking a look at my data. Always ready for a new problem, he sits right down and begins to do some quick calculations and to draw a few rough plots. Then he writes out a "Rasch" model for rating scales and we try to apply it to my data by hand. It is May 1960.

Georg's new model makes its public debut in his June 1960 Berkeley Symposium talk and travels home to Denmark to become the basis for Erling Andersen's education. We never finish applying it by hand but after I spend the spring of 1964 and, then with Bruce Choppin, the summer of 1965 in Copenhagen with Georg the new model finally gets applied to my semantic differential data through a pairwise FORTRAN algorithm, "BIGPAR," written by Bruce in the fall of 1965.

The day after my family and I arrived in Copenhagen in May 1964, I went to Georg's Institute about 11am. He was very happy to see me, showed me around quickly and hurried me off to lunch at his "nearest favorite restaurant," The Little Prince, where Georg was very well known to the management. Course by course, the proprietor brought us samples of every kind of dish imaginable. In Denmark they call this the "Alretning" which I believe means "everything in the kitchen." Georg encouraged me with "the advice the wise old Chinaman gave his son. If one eats slowly there is no limit to how much one can eat."

So we ate slowly and for hours. Frequently in the course of our infinite banquet we stopped religiously to toast one another and slake our thirst. This was especially important when eating herrings on black bread smeared with lard — a Danish delicacy.

After each bite it was de rigueur to look directly into one another's eyes, raise our glasses toward each other, emit a hearty "Skol" and down the 2 ounces of Akvavit in a gulp. This was necessary so that "the herring could swim." Two ounces of liquid, however, almost always proved insufficient to keep the herrings happy. So we usually followed the Akvavit with a half bottle of good Danish beer "to keep those herrings swimming."

Later, as we moved on from fish to beef, we shifted naturally to a "nice red wine" which kept us and I suppose the herrings swimming through meat and cheese but had to yield to an even "nicer white wine" to float fruits and desserts which in their turn must be saluted farewell with some "fine cognac." The proprietor who had been with us off and on all afternoon finally sat down with us at about 3:30 to help smoke a rich cigar and sip "very old Madeira." Georg apologized that he himself had never learned to smoke. But he assured us that his dear wife Nille did smoke and especially liked cigars.

Most of the time we did not meet at his Institute. Instead I took a perfect commuter train out to suburban Holte where he lived in a handsome mansion of many large rooms. Our mathematical work, however, took place upstairs in a rather small bedroom because that was the only place in the house where Nille had allowed Georg to install a blackboard. And without a blackboard, Georg could not work at all.

Georg had a regular round of consultations at various research institutes: The Army, The Serum Institute, The Eugenics Society, and Erik Thomsen's Institute for Educational Research which published Georg's great book.

These consultations usually took place after lunch. Georg would introduce me to everyone there, settle down in the big chair and invite the young men attending to report their progress with the measurement research they were doing under his direction. Once they got started Georg's eyes would fall shut and it would look for all the world like he was definitely sound asleep. Not at all surprising considering what we had had for lunch. This usually embarrassed the host who would hasten to my side and whisper into my ear that Georg was not really asleep. And perhaps not. For when the reports where done and the voices of the young reporters faded away, Georg would shake himself, open his eyes, tell them in detail exactly what to do next and rush us off to the next consultation.

When Georg and Nille gave us a dinner party out in Holte. Georg met each guest at the door, asked them what they would like to drink, and then, whatever they asked for, be it sherry, whiskey, vermouth or a dry martini, always had their first drink with them. He had a vodka martini with Claire and then a Bourbon whiskey with me.

At the dinner table a large bottle of red wine was put between every lady and gentleman so that the gentleman could keep the lady's glass full without inflicting upon her the embarrassment of asking for more. Throughout the many courses, whenever a guest caught the eye of another anywhere around the table, each grasped their wine glass firmly, raised it high, invoked a hearty "Skol" and finished the glass. As far as I can remember it was a lovely evening. I believe that was the evening Nille taught me to whisper endearments in Danish into the inviting ears of her beautiful daughters.

My subsequent gastronomical adventures with Georg never fell short of our first lunch. On Laesoe in August 1967 where I spent a month in his 200-year-old thatched roof farm house, we began each day by cooking a fine English breakfast which we served to Nille on a tray in her bedroom and then downed ourselves in the little dining room that looked out on the yard.

Then Georg would take me back to his office/bedroom, "created out of the former pigpen of the farm house" where one wall was a large blackboard. There we would spend 3 or 4 hours working on the mathematics and implications of his measurement models and would just be getting really serious when the sound of clinking glasses would drift down the garden path toward our mathematical sanctuary. It was Nille with a choice of cocktails before lunch.

Whenever possible lunch was in the garden and it was always fulsome: herrings, cheeses, cold meats and salads, and, of course, the essential Akvavit and beer to help it down. Needless to say, after lunch we all napped or perhaps "passed out" would be a better description.

About 3pm Georg would push his head through the small window just above the bed in my tiny guest room, look fiercely down upon my unconscious form and shout, "BOO!!" That was when we submitted ourselves to Nille's devotion to race car driving and surged out to explore the island. Georg always sat in back, clutching the dog, "just in case." We careened around the narrow lanes of the little flat island to visit Nille's many island conquests, the fishing folk who lived on the island for whom Nille was the grandest of urban ladies.

We usually took a large box of candy to the island "Fat Lady," so fat in fact that she had not been able to squeeze through a door or window of the room she inhabited for decades. The "Fat Lady" held court every afternoon, listened to and resolved family and financial disagreements, and told fortunes. The grateful islanders never failed to bring her a few more pieces of candy.

When the weather was warm we went to Danzigmann beach, a sandy peninsula jutting out into the Kattegat toward Sweden. We changed into our bathing suits in front of one another without the least self-consciousness. Georg, who was then 66, set off on his "traditional" run way down the beach and back and then we threw ourselves into the 50-degree water for a brief and extraordinarily invigorating "swim." Georg usually did more of that than I did. Nille took the sun. And then of course we had a "bite to eat" which often took the better part of an hour to complete.

In the evening a local lady referred to by Nille as "Mrs. Laeso" served by candlelight the sumptuous banquet that Nille had somehow gathered and supervised during the morning and perhaps when Georg and I were napping.

There were many courses and several wines. Often there was amazing, "just caught today," fish, virgin lobsters, and crabs which Nille had collected from her fishing friends down at the docks. As the evening darkened we talked about old times, their childhoods, their young marriage, the hardships of the 1930's and the war. Often as Georg remembered a particular time or moment he wept with the joy and sadness of it.

After the long meal we usually went into the next room, put an Italian or French opera on the phonograph, sipped cognac and/or Madeira, and Nille and I smoked our cigars.

Once in a while we drove out into the Laesoe night to visit a party at a fisherman's thatched cottage. The light was by candle as no electricity had as yet reached these cottages. The music was homemade and the dancing lively, much like our American folk dancing. Most of the dancing was done by the women, as the men seemed cautious about becoming involved in anything so impulsive. Nille sported about the room arousing excited, happy conversation with the ladies, introducing me each time, and then getting me to dance with each lady in turn and also having a dance or two herself. All the while Georg would sit contentedly in a comfortable corner sipping beer. "On occasions like these, I only get involved at the highest diplomatic level."

Most nights before we finally retired we took Nille's dog for a walk down the country road beyond the cottage. Sometimes it was pitch black, sometimes bright moonlight. We held hands and talked and laughed as we walked.

I worked and played with Georg for 20 years. He was

always happy, optimistic, full of fun, ready for anything. He loved puns and knew countless anecdotes of endearing human foibles. He was generous, wise, infinitely forgiving, and the most modest genius I have ever met.

Andersen, E.B. 1973. Conditional Inference and Models for Measuring. Copenhagen: Mentalhygiejnisk Forlag. 1973 b. A goodness of fit test for the Rasch model. Psychometrika 3 8: 123-40.

Andersen, E.B. 1977. Sufficient statistics and latent trait models. Psychometrika 42: 69-81.

Loevinger, J. 1947. A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs 61.

Loevinger, J. 1965. Person and population as psychometric concepts. Psychological Review 72: 143-55.

Rasch, G. 1923. Notes on the equations of Lagrange (with N. Nielsen). Det. kgl. Danske videnskabernes selskab. Mathematisk-fysiske meddelelsev 5, no. 7: 1-24.

1934. On Matrix Algebra and Its Application to Difference and Differential Equations. Copenhagen.

1934. On the reabsorption of cerebrospinal fluid (with M. Fog and G. Stürup). Skandinavischen Archiv für Physiologie 69: 127-50.

1947a. Recent biometrics developments in Denmark. Biometrics 4: 172-75.

1947b. On the evaluation of intelligence tests. Kobenhavns Universitets psykologiske Laboratorium.

1948. A functional equation for Wishart's distribution. Annals of Mathematical Statistics 19: 262-66.

1953. On simultaneous factor analysis in several populations. Uppsala Symposium on Psychological Factor Analysis. Nordisk Psykologi's Monograph Series 3: 65-71, 76-79, 82-88, 90. Uppsala.

1960. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research.

1961. On general laws and meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 4: 321-33. Berkeley: University of California Press.

1966. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology 19: 49-57.

1967. An informal report on the present state of a theory of objectivity in comparisons. In Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof." L. J. van der Kamp and C. A. J. Viek, eds. Leiden.

1968. A mathematical theory of objectivity and its consequences for model construction. In Report from European Meeting on Statistics, Econometrics and Management Sciences. Amsterdam.

1969. Models for description of the time-space distribution of traffic accidents. Symposium on the Use of Statistical Methods in the Analysis of Road Accidents. Organization for Economic Cooperation and Development Report No. 9.

1972. Objektvitet i samfundsvidenskaberue et metodeproblem. Nationalekonomisk Tidsskrift 110: 161-96.

1977. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy 14: 58-94.

The Foreword by Benjamin D. Wright to Georg Rasch's "Probabilistic Models for Some Intelligence and Attainment Tests", Chicago: University of Chicago Press, 1980; MESA Press, 1992.

B. F. Skinner, A Case History in Scientific Method. The American Psychologist 11 (1956), p. 221-33.

J. Zubin et al., Experimental Abnormal Psychology. Columbia University Store. New York 1955. Mimeographed. - p. 2-28.

The Preface by Georg Rasch to his "Probabilistic Models for Some Intelligence and Attainment Tests", Chicago: University of Chicago Press, 1980; MESA Press, 1992.

[1] The quotations come from David Andrich's interviews with Rasch on Laesoe June 1979 and Rasch's February 1980 letter to me.

Ben Wright: the Measure of the Man

John Michael Linacre, Ph.D.

Benjamin D. Wright

For over 30 years Benjamin D. (Ben) Wright has been the leading expert on measurement (in its usual sense) in the social sciences. For many years he was its only conspicuous proponent. Yet this inevitable situation came about seemingly by accident.

Ben was raised in pre-WWII New York City. His mother was a Professor of Psychology at New York University, but the exciting field of study was physics. Quantum mechanics seemed to be the key to an intriguing, dynamic future. WWII was joined, and in 1944 Ben volunteered for the Navy. As part of his training to become a Naval officer, he was sent to Cornell University where he obtained a Bachelor's degree with honors in Physics and Philosophy.

The war concluded, and Ben embarked, not onboard ship, but on his intended career in advanced physics. In 1947 he took a job at the Bell Telephone Laboratories in New Jersey to work with Charles H. Townes on microwave spectroscopy. Then, in 1948, he became a research assistant to Prof. Robert S. Mulliken at the University of Chicago to work on ultraviolet absorption spectra. His research entailed performing the same experiment over and over again. Each experiment required many precise measurements. Almost all experiments ended up invalid. There were incorrect experimental conditions, flawed experimental procedures, and human errors. Finally an experiment yielded results that documented theoretical predictions in a useful way. That experiment would be deemed a success, and the next experiment would commence. This research was ideal for obsessive introverts. But, despite his love for physics, Ben was not one of those. So he looked around for a more lively field of study. His first choice was English, but the English professor he interviewed was so unhappy with his life that Ben looked further.

Society at large was just becoming aware of the problem of the mentally disordered. It was still routine to incarcerate such people in a lunatic asylum. For mentally disturbed children, this implied a life sentence. Bruno Bettelheim had a broader vision. He was convinced that seriously disturbed children could be helped to live productive lives at some level. He took on the Orthogenic School and engaged on a radical and highly experimental program to discover how to help children whom others had rejected as beyond help. Ben was fascinated by this daunting challenge, and so, in 1950, he joined the School as a counselor of schizophrenic children and Bruno's research assistant. In later years, Bruno was criticized for his many failures, but Ben already knew from his experience in physics that it is the long road of learned-from failure that leads to success.

Ben now embarked on the study of Freudian psychoanalysis and psychotherapy, but maintained his interest in mathematics and measurement. He published two papers with Bruno (1955, 1957) focusing on teachers and counselors, rather than children. But ultimately the emotional, mental, and even physical stress of dealing with dysfunctional children became overwhelming. Ben began to realize that child psychoanalysis might not be the way for him after all.

Bruno was a Professor in the Department of Education at the University of Chicago. The Department encountered a sudden need for an instructor in introductory statistics and Bruno nominated Ben because of his ease around numbers. So Ben started teaching statistics in 1956, but soon ran into trouble. He noticed that the statistical textbook gave errone-

ous advice. Accordingly Ben followed his training in physics, and started teaching according to his theory of statistics rather than parroting accepted wisdom. This soon drew the ire of the Education faculty as they encountered students who had not been indoctrinated into the conventional statistical lore. The Chair of the Department, Frank Chase, supported Ben, but the matter was finally brought before the University's foremost statistician, Prof. Leonard "Jimmy" Savage. Jimmy discerned that Ben was indeed correct. Thus Ben's status as a maverick statistician was confirmed.

Louis Thurstone had been active in the University's Psychology Department, advocating the theory and practice of factor analysis until 1950, and Ben had gotten to know him in 1948. In 1959, Ben took advantage of the University's recently acquired Univac I (1 kilobyte) computer to write a factor analysis program. This was part of "exploratory work on ways to convert observational and test data to meaningful measures" (Orden, 1961, p.11). Over the next few years, Ben performed hundreds of analyses for clients, using the resulting income to support his wife, Claire, and children Amy, Sara, Chris, and Andy. The clients, however, were frustrated. Factor analysis proved to be highly sample- and analyst-dependent. Each new sample of the "same" data yielded a different factor structure. Factor analysis was clearly not the road to scientific progress.

In 1959, Jimmy Savage ran into Danish mathematician Georg Rasch at a Biometrics Society meeting in Washington D.C. (Georg was a founding member). Jimmy had gotten to know Georg in the Autumn of 1947 when Georg was a guest at the Cowles Commission for Research in Economics at the University of Chicago. Rasch had also published papers on factor analysis (1953), but it was the need to tell the world of his recent discoveries in social science measurement that Georg impressed upon Jimmy.

Shortly afterwards, Jimmy talked about Georg's work to Ben, and Ben expressed some interest. Jimmy had funds for a visiting professorship, so he said: "Well, Ben, if you tell me to have him come, I'll bring him. I don't see a reason for the Statistics Department to have him. But, if you think the people in Psychology or Education will be interested, then I'll bring him." So Georg came to the University of Chicago in 1960, and Ben felt himself obligated to attend Georg's lectures.

Georg's first lecture was heavily attended by the Statistics Department and the statistical people in the Social Science Division. In his lecture Rasch criticized factor analysis, but, more significantly, his teaching style was bombastic and uncompromising. As the lecture series continued, people stopped coming. The social scientists couldn't understand the math. The statisticians thought he might be insulting them. Jimmy fell asleep about half way through the first lecture and slept all the way through the second. Then he stopped coming. Ben felt concerned about Georg being deserted by his audience and also discerned that what was being said was interesting. And Georg didn't give up. He brought in his notebook. He opened it carefully. He gave his lecture, even when there was no one there but Ben, his last student. So they made friends. They discussed methods to analyze Ben's semantic differential data. But then Rasch's visit was over and he went back to Denmark.

Ben and Georg maintained desultory contact over the next three years. Then, in 1964, when Ben again encountered the problem of analyzing semantic differentials, he used a visit to Georg as an excuse to take a trip overseas.

In Denmark, Georg and his wife, Nille, proved genial hosts to Ben, Claire, and their four children. Georg spent the mornings lecturing Ben on math and statistics. He rejected the conventional emphasis of social scientists on summary statistics, such as correlations and reliabilities, and went right to the observation itself and modeled it. To Ben this made sense, in fact, better sense than anything he had heard previously.

Claire and Ben

When Ben returned to Denmark in 1965, he took along graduate student Bruce Choppin. On their return to Chicago they got right to work writing FORTRAN programs for all the algorithms described in Georg's book (1960). The theory and

technique were presented at a Mid-West Psychological Association symposium, encompassing all those interested in Rasch measurement, in the fall of 1965 in Chicago. That was the debut of Rasch work in this country.

Again Ben might have drifted away from measurement, except that he encountered Nargis Panchapakesan, a physicist from Calcutta with an interest in education. Ben persuaded her to get another Ph.D. while developing effective estimation procedures for Rasch measures. By 1967, the work of Ben, Bruce, and Nargis was bearing fruit in the newly implemented UCON estimation procedure. In the spring, Ben presented a paper to the Psychometric Society. Then, in the fall, Benjamin Bloom, at Georg Rasch's instigation, invited Ben to speak at the 1967 Educational Testing Service (ETS) Invitational Conference. Ben felt that ETS talk, and a few published papers, would surely allow him to pass the baton to other researchers and lead to the speedy completion of his own Rasch work. But this was not to be.

The introduction of the UCON procedure marked a change in the relationship between Georg and Ben. This new development, which melded theoretical ideal with practical necessity, offended Rasch's mathematical sensibility. Ben was no longer Georg's compliant disciple, but was becoming an authority in his own right.

The practical application of Rasch measurement now proceeded apace. The first ever pre-session at the Annual meeting of the American Educational Research Association (1969) was on the Rasch model. Researchers, in small numbers at first, started to become interested in capitalizing on Rasch measurement. The second pre-session in 1970 proved to be the start of the long-continuing Rasch-based testing program operated by the Portland (Oregon) Public Schools. And then students started to enroll with Ben. David A. Andrich in 1971 (Ph.D. '73), Graham A. Douglas in 1972 (Ph.D. '75), Geofferey N. Masters in 1977 (Ph.D. '80), and the flow continues. Ben has chaired 110 Ph.D. committees, of which 75 focused on Rasch measurement.

Ben is as active as ever, in many ways more so. His experience in constructing Rasch measures from many different types of social science data, and in reviewing analyses performed by others, far exceeds that of any other researcher. Yet he perceives there are vast areas of observational data waiting to be addressed and still much to be learned. Rasch measurement has now become exactly that intriguing and dynamic field of study for which Ben yearned as a young man.

Bettelheim B. & Wright B.D. 1955. Staff Development in a Treatment Institution. *The American Journal of Orthopsychiatry*, Vol. XXV, No. 4, October p. 705-719.

Orden A. (1961) A Report on Use of the UNIVAC, 1958-1961. University of Chicago, Operations Analysis Laboratory.

Rasch G. 1953. On simultaneous factor analysis in several populations. Uppsala Symposium on Psychological Factor Analysis. *Nordisk Psykologi's Monograph Series* 3:65-71, 76-79, 82-88, 90.

Wright B.D. 1967. Rasch Model for Item Analysis. Psychometric Society Annual Meeting. University of Wisconsin. Madison.

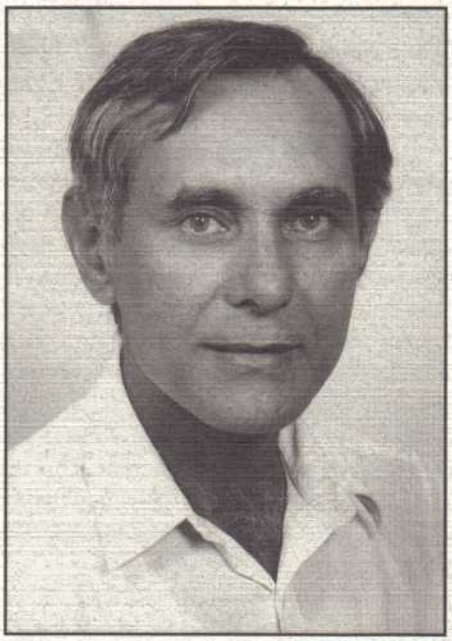
Wright B.D. 1967. Sample-free Test Calibration and Person Measurement. Published in Bloom B. (Ed.) *Proceedings of the ETS 1967 Invitational Conference on Testing Problems* (1968), 85-101.

Wright B.D. & Bettelheim B. 1957. Professional Identity and Personal Rewards in Teaching. *The Elementary School Journal*. March p. 297-307

"Measurement is not just any arbitrary arithmetical manipulation of responses; it is a theory of the phenomenon being measured. The theory may be relatively strong or weak in the assumptions it makes, but theoretical assumptions are being made, implicitly or explicitly. If the theory is wrong, or if our numerical relational system is not related homomorphically to our empirical relational system, then the arithmetic we perform on our numbers will not answer questions about the empirical relational system — or worse, it will answer them incorrectly. In a time that permits the facile use of various computer algorithms to turn responses into numbers, this point is important. All measurement is theory in some sense and we ignore the theory at our peril." pp. 245-6 in Anderson, A. B., Basilevsky, Alexander, Hum, Derek P. J. Measurement: Theory and Techniques. In Peter H. Rossi, James D. Wright & Andy B. Anderson (Eds.) 1983. Handbook of Survey Research. San Diego, CA: Academic Press, pp. 231-287.

David Andrich:
A Genius From Down Under

Linda Webster, Ph.D.



David Andrich

David Andrich has accomplished more in his first fifty-seven years than most expert teams might hope to accomplish in several lifetimes. From research grants to fellowships to award-winning research, Andrich is worthy of note in the field of measurement.

David Andrich was born in Perth, Western Australia in 1941. Earning a Bachelor of Science degree in Mathematics from The University of Western Australia in 1961, he began teaching math while working on additional curricula. In 1968, he completed a Bachelor of Education Degree with First Class Honors and was awarded the Cameron Prize for his B.Ed. Honors Thesis.

Three years later, in 1971, he completed an M.Ed. at Western Australia and headed for the United States as a Fulbright Fellow. Over the next two years, 1971-1973, Andrich completed the Ph.D. at the University of Chicago and earned the Susan Colver Rosenberger Prize for the best Ph.D. thesis in the Division of Social Sciences. His internationally known Ph.D. committee included psychoanalyst and physicist B.D. Wright, statistician Shelby Haberman, and R. D. Bock, a quantitative psychologist.

In 1977, Andrich spent six months working with Danish mathematician, Georg Rasch, at the Danish Institute for Education Research. He has twice been invited to six-month stints at the University of Chicago in 1977 and 1986. The University of Trento, Italy, appointed him Visiting Professor in 1991 for two months and in 1993 for three months. He has given lectures or workshops in Great Britain, Denmark, Germany, Austria, The Netherlands, Italy, Hong Kong, Singapore, the United States, Canada, and Australia.

His "day" job has kept him rather busy, as well. In 1985, he was appointed Professor of Education at Murdoch University in Western Australia, and he held the position of Dean from 1988 through 1990.

Elected a Fellow of the Academy of Social Sciences of Australia in 1990, Andrich has an extensive research resume to support that election. In 1989, he completed a commissioned study of Upper Secondary School Certification and Tertiary Entrance for the Minister of Education in Western Australia. Last year, he completed the International Perspectives on Selection Methods of Entry into Higher Education for the Higher Education Council, DEETYA. He serves on the editorial boards of Psychometrika, Applied Psychological Measurement, the Journal of Educational Measurement, the Australian Journal of Education, and Education Research and Perspectives.

In 1988, Sage Publications brought out his Rasch Models for Measurement under the Quantitative Applications in the Social Sciences Series. Andrich has contributed chapters to a variety of books and conference proceedings, as well as writing articles for eight major journals.

He has been awarded research grants by the Australian Research Council beginning in 1985 and has completed research on the Intellectual Development of Pre-Adolescent and Adolescent Children, Advancing Psychometric Theory for Studying Profiles of Performance and Structuring and Assessing the Latitude of Acceptance.

His professional memberships and appointments are international in scope. These include the American Educational Research Association, Research Methodology Chapter of the International Sociological Association, the Australian Association for Research in Education, and the National Council for Measurement in Education. He served on the National Consultative Council of the Federal Government from 1989-1991.

His current research includes the affective development and assessment of opinion, attitude and preference and choice, educational and social processes including assessment and selection, integrating qualitative and quantitative methods in the social sciences, the development of models for the measurement of processes in the social sciences, and the philosophy of social measurement. He is known for his innovative work in modern test theory and, in particular, Rasch models for measurement.

Dr. Linda J. Webster is an associate professor of speech and director of the university Honors Program at the University of Arkansas, Monticello. She was installed as first vice-president of the Arkansas State Communication Association on March 6 and will be assuming the office of president in 1999. She is the editor of the "Journal of Communication Studies" and is a practicing journalist who teaches introductory newswriting along with courses in rhetoric, women's studies, museum management, and interdisciplinary studies in the Honors Program.

Some Insights into Objective Measurement

David Andrich, Ph.D.

There are many aspects of my studies in quantitative methods with Georg Rasch that have been important in my work, but I will focus here on two related insights that were the most telling in affecting my outlook on measurement.

I was a student at the University of Chicago in 1971 – 1973, at a time when the Department was extremely exciting. Among his many favors, Ben Wright did his biggest one by introducing me to Georg Rasch. On completing my Ph.D., I visited Rasch in Denmark and arranged for him to be a visiting professor in the Departments of Education and Mathematics at the University of Western Australia in 1974. I spent many hours during the day with Georg, and my wife Joan and I enjoyed the company of Georg and his wife Nille during the evenings and weekends for seven months. We then repeated the pleasure in Denmark in 1977 for another five months.

In studying general quantitative methods in the social sciences, I had learned a whole range of techniques and skills for using models and analyzing data. In addition, however, I learned the implied general philosophical position behind these studies, namely, that the task is to find a model that accounts for the data. One could debate this position in general, but in the case of measurement, I realized through the work with Rasch that the case for his class of models does not depend on modeling any particular data. This was a very important shift in perspective for me, and I believe that where there is controversy in the use of Rasch models, it is where people consider that the choice of one model or another rests essentially on how the models account for data. The Rasch class of models are justified as expressions of the requirements of measurements; they are not justified as descriptions of data. Although it now seems obvious, at the time it seemed a very important insight to me.

The case for the model rests on the requirements of measurement, and if data are to be transformed to measurements, then they must be valid expressions of the construct in all the traditional senses, and in addition, need to meet the requirements of the Rasch class of models. In the special case of dichotomous responses, the discrimination at the difficulties (item thresholds) has to equal. To estimate the discriminations destroys the requirement of invariance of person ability estimates in the model, and if items have different discriminations, then item difficulties can take different orders depending on the distribution of the persons. Of course, it is an empirical question as to the degree that real data show equal discriminations, and many data sets will not immediately have equivalent discriminations. That is why it is tempting to try to estimate the discriminations. It is not surprising that data sets which have been collected without an eye to these requirements do not follow a model with the equal discriminations criterion — indeed it is surprising how many data sets do follow the models sufficiently well to be informative.

My second insight came with the resolution of the coefficients in Rasch's form of the multicategory model. Rasch generalized his dichotomous model to one for many categories as a multidimensional model and then specialized it to the case of a single location parameter for the persons and items. In the process he had a coefficient and a scoring function for each category.

These coefficients and scoring functions were difficult to make sense of in any concrete way. I constructed the model for the response of a single person to a single item beginning with the simple model for dichotomous responses at each threshold of a multicategory item, and that gave me the integer scoring function and the resolution of the category coefficients into the sums of successive thresholds. I did this while Rasch was in Perth in 1974. However, more data sets than not showed reversed thresholds in their estimates, which was inconsistent with the construction of the model. While in Copenhagen, in 1977, Erling Andersen showed me a prepublication copy of a paper to appear in Psychometrika, in which he showed that the scoring functions had to have a constraint. My integer scoring functions had such a constraint, which confirmed to me that I was on the right track, but as I indicated, more data sets than not showed a problem with the estimates. My insight came in realizing that when the threshold estimates were reversed, this was not a problem with the model, but with the data. In particular, if discriminations at the thresholds were not equal, then it was possible, as in the dichotomous case, to get any ordering of the thresholds, depending on the distribution of the persons. Again, after the formalization, the result seemed obvious and a simple generalization of the dichotomous case. However, at the time, it revealed the level of resistance in my mind in taking seriously that the case for the model rests on criteria independent of the data, and not in modeling data. Because of our traditional studies in quantitative methods, it is much easier to think that the model should describe whatever data are at hand, and it is difficult to maintain in our thinking that the case for the Rasch models become independent of data. It is also difficult to resist the temptation to use other models to model the data, rather than to examine the data to see how and why they violate the requirements of measurement.

Methodology and Morality

William P. Fisher, Jr., Ph.D.

Are qualitative (feminine, collaborative) and quantitative (masculine, domineering) methodologies merely better or worse ways of addressing different problems (Short-DeGraff & A. Fisher, RMT 7:3, p.301) or do they connote different moralities?

Even in measurement, it is clear that competing methodologies reflect rival systems of ethics. Richard Jaeger, in his 1987 NCME Presidential Address, quoted Wright (1977, p.77) "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." In contrast, Jaeger quotes Lindquist (1953, p.35) "The objective [of an educational test] is handed down by those agents of society who are responsible for decisions concerning educational objectives, and what the test constructor must do is to attempt to incorporate that definition as clearly and exactly as possible in the examination that he builds."

Notice that in Wright's approach, the community of objective-definers, test-constructors, and tested-persons is egalitarian. Every member of the community has a voice in deciding which items are useful and which are not. The basic ethic includes fair play, justice, and democracy — and even aesthetics, as represented by the mathematical elegance of the Rasch model.

Lindquist, however, is concerned with content validity rather than construct validity. There is elitist and centralized control of the objective. Test-constructors and tested-persons are at the mercy of the test-definers. "The definition of the objective is sacrosanct" (Lindquist ibid.).

Here the Rasch debate is but a microcosm of the qualitative/quantitative debate, since virtually all quantitative methods proceed in a manner more akin to Lindquist than to Wright. "The question is not about how to define words like *truth* or *rationality* or *knowledge* or *philosophy*, but about what self-image our society should have of itself" (Rorty, 1985, p.11). Now, as much as in Galileo's time, our scientific methodology reflects our innermost selves.

The appeal of Rasch methodology is not in its somewhat abstract scientific qualities, but in its capacity to build solidarity and community by ensuring that everyone can contribute in a constructive way, building consensus while simultaneously acknowledging and learning from dissent. High quality measuring instruments extend our conversation into new domains, justifying our theoretical constructs and their measurement, not through appeal to an arbitrary, unfeeling higher authority, but through the way they emerge from within the community affected by them. The mutual interaction of subject and object is unavoidable. Rasch helps us to capitalize on this mutual interaction and so increase its flow.

E.F. Lindquist (1953) Selecting appropriate score scales for tests (Discussion), Proc. 1952 Invit. Conf. on Testing Problems. Princeton, NJ: ETS

R. Rorty (1985) Solidarity or objectivity. In J. Rajchman & C. West (Eds.) Post-Analytic Philosophy. New York: Columbia

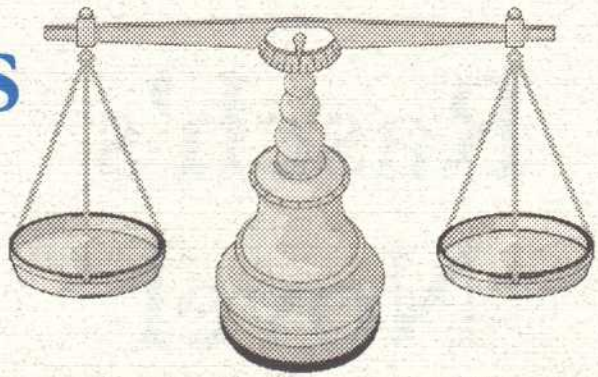
B.D. Wright (1977) Solving measurement problems with the Rasch model. Jou Ed Meas 14,2 p.97-116.

Theory vs. Practice

"People sometimes say, *This is right in theory but it doesn't work in practice.* They ought to say, *This is wrong in theory and consequently it is wrong in practice.* There is no true theory which could be wrong in practice. This contrast between theory and practice is contrived by people who want to escape hard and thorough thinking. They like to abide in the shallowness of accustomed practices, on the surface of a so-called *experience*. They will accept nothing but a repeated confirmation of something they already know or believe. Only those questions for truth which have challenged and disturbed centuries of practice have brought about a fundamental transformation of practice. This is true of the history of science, morals and religion."

Paul Tillich, "Doing the Truth", in "The Shaking of the Foundations", 1949.

Rasch Invents "Ounces"



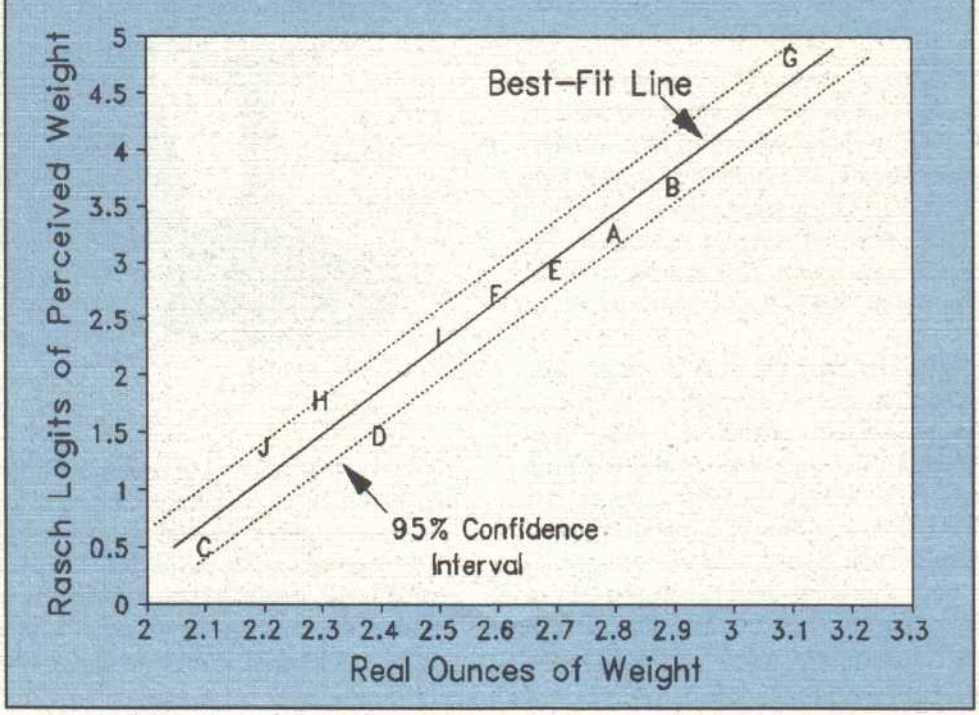
Ellie Choi, University of Chicago

One might wonder how civilization ever arrived at the efficient and reliable abstraction we call "weight" — measured out, for instance, in ounces? We could not get along without this lovely abstraction. Not only physics and engineering, but commerce also would collapse. But where did the measurement of weight come from? How did it develop?

We cannot trace its entire history, because most is unrecorded. But we can reenact, right now, an experiment which shows how it must have come about. We can demonstrate the irresistible and nearly perfect connection between the simplest possible hand-to-hand perceptual comparison and professionally measured weight. All we have to do is to compare pairs of objects for their apparent relative heft by passing them back and forth from hand to hand and record which one seems heavier. Nothing more exact or demanding is needed. An entirely psychometric, that is mathematical, construction built from a collection of these simplest of all observations produces a linear equivalent to "objective" weight.

Ellie Choi poured different amounts of rice into 10 unmarked paper cups, sealed the cups, and labeled them "A" through "J" at random. Then she asked each of 13 students she happened to encounter to pick up pairs of these cups, one in each hand, pass them back and forth and then tell her which cup seemed heavier. The 10 cups produced 45 pairings per student. Her experiment produced 580 separate paired comparisons with the heavier-feeling cup scored "1" and the lighter-feeling cup scored "0" each time. After Ellie had collected these data, she weighed each cup on a postal meter to determine its "official" weight in ounces.

Figure. Perceived and real weights of 10 cups of rice.



When Ellie analyzed these simple dichotomous paired comparisons with the Rasch measurement program *Facets*, she found that the Rasch calibrations of the 10 cups in logits formed a statistically linear relation with their weight in ounces. Here is her picture of this relationship.

The implications of Ellie's experiment for the history of measurement is that the linear abstraction of "weight" has been resident in our simplest perceptual judgements since our beginning, whenever that was. All we did over all those centuries was to discover, step-by-step, how to make the implications of what we felt in our hands into objective, reproducible measures.

Ellie's experiment was replicated by Natalie Colabianchi last autumn with the same result.

Choi, Sungeon "Ellie" (1995) Using Paired Comparisons to Determine Weight Perception. Unpublished paper. University of Chicago.

Rasch's Novel Wisdom

William P. Fisher, Ph.D.

The wisdom of Rasch measurement is the same as what Milan Kundera calls the "wisdom of the novel ... the fascinating imaginative realm where no one owns the truth and everyone has the right to be understood" (quoted by Richard Rorty in the April 1994 University of Chicago Magazine, p. 23). Data are fictions that provide evidence in support of measures, but do not determine or prove them; thus, to the extent that they enable learning about people and ideas, data are fascinating imaginative realms.

In these realms, no one owns the truth about the variable or the measures; neither the test designer nor the person with the highest measure can know all there is to know about the variable. The variable and the measures always remain open to new interpretations; indeed, the notion of scale-free measurement could be taken to follow from the observation that the universe of items embodying a particular construct is potentially infinite. Since it is impractical to consider administering or even conceiving all of the items on a variable, it becomes necessary to recognize each individual instrument as just one interpretation of the variable, and to evaluate it in terms of its targeting and the consistency of the data it produces.

Finally, data consistency is the means by which everyone's right to be understood is realized. Fit statistics (measures of data consistency) inform us about the validity of a measure, and indicate when someone's ability or attitude is likely to be misrepresented by a measure.

So interpretation of a Rasch analysis is a matter of asking what story is told by the data. What's the plot? Who are the characters? What's the setting? Is there a subplot? Is there a single overall theme, or does the story try to go in too many directions at once?

In actual practice, of course, most tests are taken to define the variable in and of themselves, so no imaginative realm is opened up. Test content is considered sacrosanct and of predetermined validity, so the truth of a test is owned by whomever wrote it, published it, or set up the learning objectives. These same people are the only ones who have the right to be understood, since they set the rules, referee, and control the scoreboard.

Rasch analysis in itself does not open up fascinating imaginative realms, unlock the truth, and give everyone the right to be understood. These things can happen in educational and psychological measurement without any help from Rasch. Conversely, Rasch's models can be applied in ways that will never free the imagination, truth, or other people. But Rasch's approach to measurement does offer us just that combination of features and techniques sufficient to the job, and it offers them to us efficiently packaged, making Rasch measurement simple, elegant, and parsimonious. Where Rasch analysis coincides with the wisdom of the novel, contemporary problems become more manageable and hope for the future is born.

Three Stages of Construct Definition

The development of construct definition follows a process that is articulated by its source of knowledge.

A. Jackson Stenner, Ph.D. & Ivan Horabin, Ph.D.

Stage 1) Instrument calibration based on personal knowledge, intuition, and subjective analysis.

Pre-Galilean discussions of temperature measurement are interspersed with references to subjective "scales" of measurement anchored by terms like "as cold as when it snows" or "too hot to touch." A recent example is the attempt to measure "health risks of exposure to ionizing radiation." The observation (quantity of ionizing radiation) is converted into a measure (health risk) via calibrations based on the observer's value system. Objective measurement of constructs in their formative stages is difficult because theory is weak.

Stage 2) Data-based instrument calibration.

17th Century temperature measurement employed data-based calibration. In Europe, two dozen "scales" competed for favor. Calibrations of thermometers were done on an instrument-by-instrument basis in the laboratory of the instrument maker. The particular readings of the thermometer, when exposed to states with known temperatures (e.g., human temperature), were used to calibrate each thermometer as it was manufactured. Measures from the same instrument maker were consistent and "specifically objective," i.e., two instruments from the same maker produced basically the same numbers. Measures from thermometers built by different instrument makers differed, and there was no common frame of reference to permit a measure's reexpression in another metric.

A recent example of second stage construct definition is "mathematics achievement." Numerous instruments (tests) exist for measuring "mathematics ability," each with its own scale. Fifty years of factor-analytic research imply that all instruments measure something in common, but there is no shared framework that permits reexpressing one measure (e.g., NAEP) in terms of another (e.g., CAT). The confusion produced by multiple metrics contributes to the lack of consensus about what is, or should be, measured under the label of "mathematics ability."

Stage 3) Theory-based instrument calibration.

Thermometers made today are manufactured and shipped to customers without reference to data on the performance characteristics of the particular instrument. Instrument calibration is accomplished via theory-based equations and tables. Manufacturing proceeds with total reliance on theory. Theory enables any measure to be reexpressed in the metric of another instrument maker (e.g., Celsius to Fahrenheit). Measures calibrated by theory are "generally objective." Any two observers given the same observation (volume displacement of mercury in a tube) will report back the same number as a measure.

The only behavioral science construct that approaches third stage development is "reading comprehension." This is because the Lexile Framework enables generally objective, theory-based measurement of reading comprehension. Reading comprehension tests can be calibrated on the same metric, without reference to the performance of actual readers. The only reference required is the Lexile equation.

A. Jackson Stenner & Ivan Horabin
28 Stoneridge Circle, Durham, NC 27705

"Applying the Rasch model in test development gives us new versions of the old statistics. These new statistics contain all of the old familiar information, but in a form which solves most of the measurement problems that have always beset traditional test construction" (Wright and Stone 1979: 24).

Wright, Benjamin D. and Mark Stone. 1979. Best Test Design. Chicago: MESA Press.

Where Do Dimensions Come From?

Are "dimensions" facts of nature waiting to be discovered, or are they artifacts of our imagination waiting to be invented?

Physicists are in no doubt about how they think of the world about them: "It seems inevitable that we should speak in terms of some definite theoretical model of the world of experience. There appears, however, to be no meaning in supposing there to exist a unique final model that we are trying to discover. We construct a model, we do not discover it" (McCrea 1983, p. 211). The idea of length is a theoretical model for an attribute of an object. Length does not exist on its own in nature, we invented it because it suits our purposes.

The idea of length is operationalized by means of devices such as rulers. Rulers are always imperfect representations of the idea of length. They are inaccurate and imprecise, but we use them because they are good enough for our purposes. Of course, every length-measuring process must be regulated to insure that the resulting number fits with our idea of length. Bending, breaking, or otherwise misapplying the ruler still produces numbers, but not numbers that fit our idea of length.

Length is apparent to us because it is visible, but what about temperature? We want to think about heat in the same way we think about length, as linear quantities. But we don't see heat in this way. Consequently we convert heat to length, or length-like numbers, by thermometers. Now we can think about and manipulate temperature in just the same way that we do length. Representation of abstract ideas requires visualization: rulers meet our need to think in a well-controlled, uniform way.

Educational tests must operate in the same way, if we want to make sense of them. We use our imaginations to invent a construct, math ability, that suits our purposes. This construct is our dimension. We express it in terms of an abstract item hierarchy: addition, substraction, multiplication, division. We operationalize it in a math test. But is this dimension useful?

We discover whether our invented dimension, our construct, has any meaning and utility beyond our own imagination by looking for confirmation and contradiction of our intentions. We analyze the responses to our test. Do the item difficulties correspond with our intended hierarchy? Do individual items maintain their locations, i.e., do they fit? Are noticeably different persons positioned at noticeably different locations on the dimension, i.e., separated, in a way that suits our purposes? Contradictions and deficiencies lead us to re-express our construct and revise our operationalization of it. Perhaps it would be more useful for our purposes for addition, substraction, multiplication, and division each to have its own dimension, but that would lead to four measures. We must choose: Are four measures more useful or more confusing than one? Theory can't answer this, only practice can. If we have only one decision to make, then we want only one measure to base it on. If that one decision is actually a series of smaller decisions, then for each of those we want only one measure. Discovering that an individual's height is multidimensional with head dimension, torso dimension, and leg dimension is of no help, and so ignored for most purposes involving subject heights.

We know that perfection in conceptualization and operationalization will never be reached. A yardstick is not perfectly unidimensional, nor perfectly precise, nor perfectly accurate. But it is good enough for our purposes. So the ruler we construct from test responses falls short in just the same way. In both cases, we maintain the meaning of our dimension by careful use and maintenance of the ruler. We screen out and investigate errant measurements, misapplications, inconsistent results, and warped test instruments. We insist that only measures in useful accord with our invented dimension have the meaning we impute to the numbers. If no such measures are found, our dimension is useless, however conceptually sound it may be. If such measures are found, then they and only they suit our purposes, and the dimension is useful, however unrefined it may be.

Further thought and investigation will always reveal that our current idea of any dimension, "length," "temperature," "math ability" is deficient, and its operationalization by our ruler is defective. Progress requires that we be prepared to base our actions on what we can usefully achieve now, rather than on the perfection of the infinitely distant future.

Ben Wright

W.H. McCrea (1983) Introductory Remarks. Phil. Trans. R. Soc. Lond. A. 310:211-213.

"Flow" as a Testing Ideal

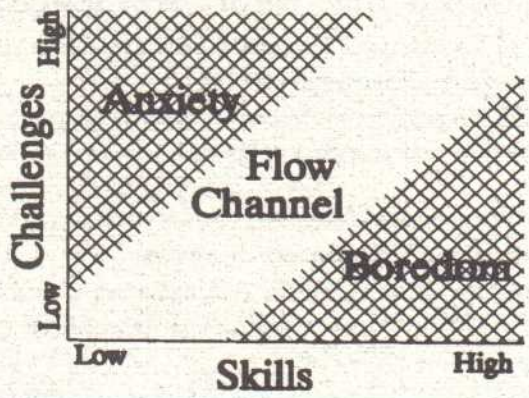
"In our studies, we found that every flow activity... provides a sense of discovery, a creative feeling of transporting the person into a new reality."

(Mihaly Csikszentmihalyi, *The Psychology of Optimal Experience*, Harper & Row, 1990 p.74)

Csikszentmihalyi describes how human activities often comprise two opposing components, which in the Diagram are characterized as Challenges and Skills. So long as the level of challenge facing the player of a game is in rough accord with the level of the player's skill, then the player will experience a "sense of discovery," or even a "previously undreamed-of state of consciousness" — that is *flow*. But as the player's skill increases, the player will grow bored. Or when the challenge of the game increases too far beyond the player's skill, frustration will set in. Both boredom and frustration inhibit the flow experience. The motivation towards enjoyment provokes one to desire to balance challenge with skill, and so to induce flow.

Tailored testing can take advantage of the phenomenon of flow to make the testing experience pleasurable and to improve individual performance. Well-targeted items will make the testing situation less irksome, perhaps even enjoyable! Targeting removes items that are too hard, so inducing anxiety, and those that are too easy, so inducing boredom. Psychometrically, the better the match between the item's difficulty and the test-taker's ability, the greater the likelihood that the situation will produce accurate measures. After a test that successfully matches item difficulties with test-taker ability, test-takers can leave feeling content that their optimum performance levels have been demonstrated, and test constructors can count on accurate measures. A flow experience for all!

Craig Deville, Ohio State University



A Savvy Test-taker.

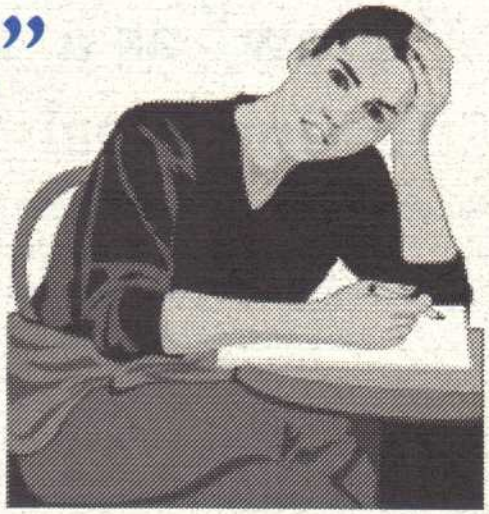
Thomas O'Neil

In 1996, I decided to sit for the GRE. Working for a company that creates and administers several computer-adaptive tests, I knew adaptive tests were usually less grueling than the pencil and paper variety. Being human, I also wondered if there was some way to exploit the adaptive nature of the test. I read in the ETS brochure that there is a minimum number of questions that had to be answered in order to receive a score. This seemed to me to be an opportunity to control when the test stopped. I decided that I would do my best on each subtest up to the minimum number of questions. Thereafter, I would take my time, endeavoring to answer all the questions correctly, but if I was uncertain of the answer, I would stop and wait for time to run out. When I actually took the test, I found myself faced with this situation only once.

I thought that this strategy would help me because all adaptive tests are based on some type of latent trait theory. Usually, the ability estimate is the log of an odds ratio (right over wrong). My thinking was that I wanted to increase the numerator (number correct) without increasing the denominator (number wrong). Provided there was no penalty after answering the designated minimum number of items and provided that I could accurately predict whether I answered a question right or wrong, this strategy should have given me a slight advantage.

As expected, I did reasonably well on the test, but I still wonder what difference the test-taking strategy really made, if any. I have run some simulations using other data sets and found some slight mean increases, but I still don't know what difference it made for me. The experience caused me to think about the role of time in a test and the treatment of incomplete test records in a variety of situations. There has been a great deal of concern about the opportunity to "cheat" on adaptive tests by manipulating responses so the algorithm is more favorable to the examinee. Having attempted a test-taking strategy that was supposed to produce more favorable results, I still have no way of knowing if it worked, or if I would have earned the same score having taken the rest of the test. Obviously, I'll never know, unless ETS asks me to take a complete test, but indicate where I would stop using this strategy and then compare the two scores. Anyhow, I prefer to think that my score is a reflection of my scholastic achievement rather than my ability to devise a "cunning" strategy.

What is the "Right" Test Length?



The "right" test length is more folklore and accident than intention. Anastasi assures us that "other things being equal, the longer a test, the more reliable it will be." Unfortunately "other things" are never equal. Nunnally mandates that for "settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated." Unfortunately he does not explain how to determine the test length that gets a .90. That's because reliability is an awkward amalgam of the length and targeting of the test, and the spread of the examinees who happen to take this test.

What's wrong with a one-item test?

1) Content Validity

To be useful a test must implement the one intended dimension. We assert our singular intention through the formulation of test items. But each item, in all its reality, inevitably invokes many dimensions. No matter how carefully constructed, the single item will be answered correctly (or incorrectly) for numerous reasons. The unidimensional intention of a test only emerges when this intention is successfully replicated by essentially identical, yet specifically unique test items. Whether an item requiring Jack and Jill to climb a hill contributes to test score as a reading, physics, or social studies item depends on the other items in the test.

2) Construct Validity

The various items in a useful test replicate our singular intention sufficiently to evoke singular manifestations we can count on to bring out the one dimension we seek to measure. Arithmetic addition is usually intended to be easier than multiplication. We could write hard multiple-digit additions that would be more difficult to answer than simple single-digit multiplications. But such a test would not realize our intention to measure increasing arithmetic skill in an orderly and easy-to-use way. Once we have successfully implemented our construct, the qualifying items define our variable, and their calibrations provide its metric benchmarks.

3) Fit

A useful test gives examinees repeated opportunity to demonstrate proficiency. An examinee may guess, make a careless error, or have unusual knowledge. One, two, or even three items provide too little evidence. We need enough replications along our one dimension to resolve any doubts about examinee performances. As doubts are resolved, the relevance of each response to our understanding of each examinee's performance becomes clear. We can focus attention on the responses that contribute to examinee measurement, reserving irrelevant responses (guesses, scanning errors, etc.) for qualitative investigation.

4) Precision

A useful test must measure precisely enough to meet its purpose. The logit precision (standard error) of an examinee's measure falls in a narrow range for a test of L items: $2/L < SEM < 3/L$. Doubling precision (halving the standard error) requires four times the items. The placement of examinee measures and confidence intervals (±SEM) on the calibrated variable shows us immediately whether the test has provided enough precision for the decisions we need to make.

When there is a criterion point, it is inevitable that some measures will be close enough (less than 2 SEM) to leave doubt whether the examinee has passed or failed. In these cases, an honest, but statistically arbitrary, pass-fail decision may have to be made. There is no statistical solution. Increasing the number of items increases test precision, but we always reach a point at which we no longer believe the added precision. If your bathroom scale reports your weight to the nearest pound, you could weigh yourself 1000 times and get an estimate of your weight to within an ounce. But you would not believe it. Your weight varies more than an ounce and, indeed, more than a pound over the course of a day.

So what is the "right" test length?

1) Enough items to clarify the test's intention and replicate out a unidimensional variable.

2) Enough person responses to each item to confirm item validity and provide a calibrated definition of the variable.

3) Enough item responses by each examinee to validate the relevance of this examinee's performance.

4) Enough responses by each examinee to enable precise-enough inferences for the decisions for which the test was constructed and administered.

Ben Wright

Cross-Language Test Equating

Richard W. Woodcock, Ph.D. and Ana Muñoz-Sandoval, Ph.D.

How can performance levels in different languages be compared? Here are the steps we used to equate four tests, each developed to measure one of four skills (vocabulary, analogies, reading, and writing) in one language (English), to four similar tests in a second language (Spanish) for the *Woodcock-Muñoz Language Surveys* (Chicago: Riverside, 1993).

Step 1. Develop a bank of English items for each of the four skill areas.

Step 2. Rasch analyze the responses of subjects (6,359, ages 2 to 95) to establish the scale underlying each of the four banks of English items. Discard misfitting items.

Step 3. Compute norm measure tables of typical English language performance at different educational levels, ages, etc.

Step 4. Select sets of items for the four published English tests from their respective banks. Compute a raw score-to-measure table for each test for field use.

Step 5. Identify a subset of items covering the full difficulty range from "easy" to "difficult" in each English test and translate into Spanish. Verify that each English item has a reasonably direct counterpart in Spanish, e.g., authority/autoridad. (Though translated from English items of known difficulty, the difficulty of the Spanish items is still unknown.)

Step 6. Develop item banks for the four Spanish tests. Imbed in each bank of Spanish items the set of equating items translated from English.

Step 7. Rasch analyze the responses of Spanish speakers from outside the USA (1,200) and Spanish speakers from within the USA (800, pre-school to university graduate). Establish item difficulties for each Spanish item bank.

Step 8. Plot the English and Spanish difficulties for equivalent items for each skill area. Do not expect perfect agreement between the two sets of calibrations, because there are language-related differences for specific item content. Evaluate outliers from the best fit line for exclusion from the set of equating items. (This step verifies success in constructing comparable English- and Spanish-language variables.)

Step 9. Select items for the published versions of the Spanish tests. Compute a raw score-to-Spanish measure table for each test.

Step 10. Compute means (M_e for English and M_s for Spanish) and standard deviations (S_e for English and S_s for Spanish) of the English and Spanish equating item calibrations for each pair of tests.

Step 11. Equate each Spanish measure, B_s, to the scale of English measures, with value B_e.

Step 12. Construct a raw score to English equated measure table for each Spanish test, so that an examiner using the Spanish tests on native Spanish-speaking subjects can use the English equated measures to identify in the English language norm tables the typical performance levels for native English speakers of equivalent language competency.

Richard W. Woodcock
University of Virginia

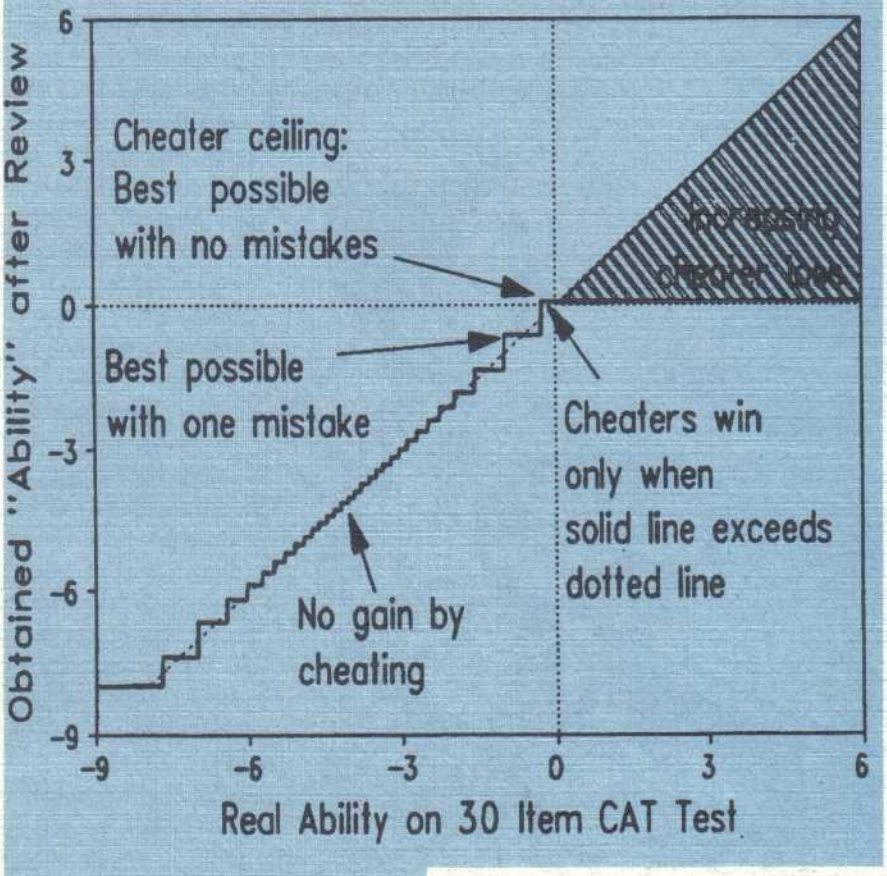
Ana Muñoz-Sandoval
Measurement/Learning/Consultants
Box 161
Tolovana Park, OR 97145

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

Rasch, 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, p. xx.

CAT and Test-Wiseness

When first introduced, MCQ tests were thought immune to test-taking strategy. We were soon disillusioned. Now computer-adaptive (CAT) tests are thought to be immune, but this time test constructors are alert. A tempting strategy is to deliberately fail the first CAT items, in order to solicit easier items from then on. This will produce an artificially high success rate and, perhaps, a higher measure than would have otherwise been obtained.



Gershon and Bergstrom (1995) considered this strategy under the best possible conditions for the potential cheater: a CAT test which allows an examinee to review and *change* any responses. This type of examinee-friendly CAT is already used in high-stakes tests and will rapidly spread, once CAT fairness becomes a priority.

Consider an extreme case in which an examinee deliberately fails all 30 items of a 30-item CAT test. After these 30 items, the algorithm would assign that examinee a minimum measure. But then, at the last moment, the examinee reviews all 30 items, most of which are very easy, and corrects all the responses. What happens?

The Plot of obtained versus real ability shows the answer. When real ability is high, all items will end up correct. But they are easy items, so the obtained ability will not be so high. Cheaters with high real abilities will invariably lose. It turns out that, at best, lower ability cheaters can obtain no more than an extra .2 logits beyond their real ability. Usually even these cheaters lose because, if they make just one slip, their obtained ability will be lower than their real ability. And now there may no longer be the opportunity to take more items to recover from that mistake, as there would be during normal CAT test administration. Should cheaters accidently exit without making corrections, they could lose 8 or more logits.

Under the most favorable circumstances this strategy can only help the examinee minutely, and even that at the risk of disaster.

A word to wise examinees:
Do not attempt this method of cheating!

Richard Gershon, Ph.D & Betty Bergstrom, Ph.D.
Computer Adaptive Technologies, Inc.

Gershon R, Bergstrom B (1995) Does cheating on CAT pay: NOT!.

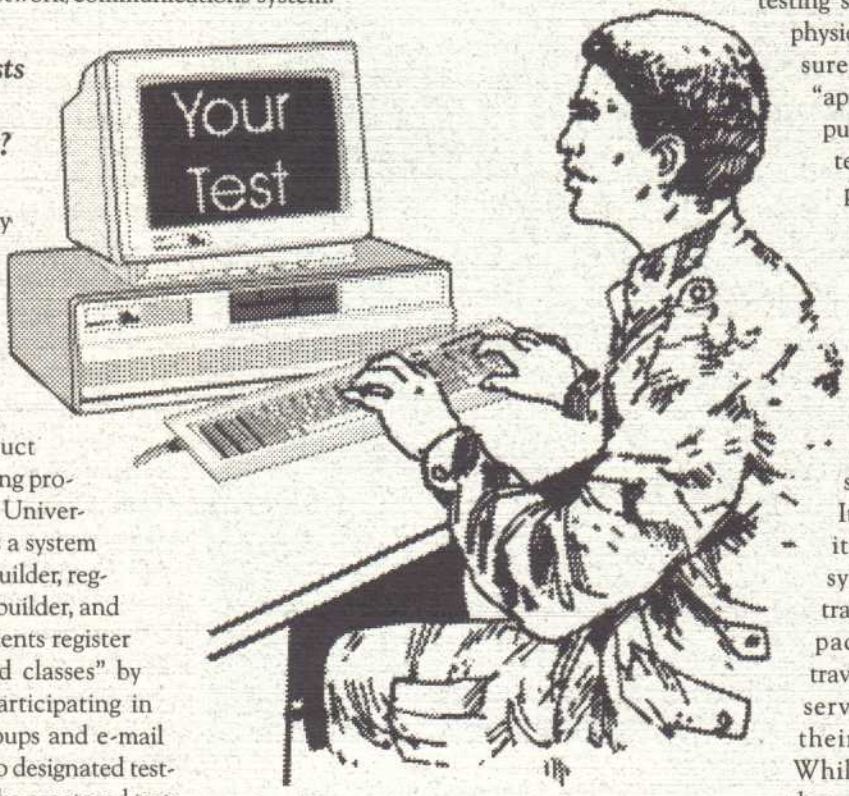
Web-Enhanced Testing

By Richard C. Gershon, Ph.D.

Q: ***What is "Web-enhanced testing?"***

A: The term "Web-enhanced testing" encompasses any aspect of testing-building, registration, delivery, administration, and scoring that is facilitated by use of the Internet, a public HTML standards-based network/communications system.

Q: ***What types of tests are appropriate for Web delivery?***



A: Technologically, any test, such as certification, performance, skill assessment, or self-evaluation, can be delivered via the Web. The University of Washington and the University of Wyoming conduct their entire distance-learning programs over the Web. "Web University," as it's called, contains a system builder, an administration builder, registration builder, a syllabus builder, and a courseware builder. Students register over the Web and "attend classes" by downloading notes and participating in listservs and discussion groups and e-mail homework. They must go to designated testing centers on campus to take proctored tests, however.

Today only non-proctored tests are appropriate for administration through the Web. At this time, identity verification is difficult through the Internet, though WebCams and retinal scanners may make this less of an issue in the future. Until then a human proctor is necessary to ensure the integrity of high-stakes tests.

Internet traffic, i.e., the volume and time differential of Web use, also impacts the type of test mounted. Low stakes or practice examinations are less likely to be overly faulted than more complex tests when the Web is running slowly. Though inconvenienced, examinees will not be severely affected.

Q: ***What about Web security?***

A: Tests delivered through the Web can have more protection than tests passed out by hand and guarded by the human eye. Paper tests must be shipped and stored in advance of a testing session and ultimately physically destroyed to ensure that copies are not "appropriated" for illicit purposes. Web-delivered tests, however, can be produced in multiple formats at the moment of distribution.

The Internet was initially designed by the military as a reasonably secure communications channel that could survive nuclear attack. Its security comes from its "packet-switching" system. The Internet transmits information in packets of bytes that travel through a number of servers before reaching their final destination. While Packet A may go through computers in Sydney, Tokyo, Moscow, and Tel Aviv before reaching its destination in Madrid, Packet B will take an entirely different route to get to the same place. Once the packet reaches its destination, the route can be traced. But the routes of subsequent packets — even from the same transmission — cannot be designated or predicted beforehand.

The security of the CATGlobal(tm) Testing Network, CAT, Inc.'s international channel of test centers, is based on this packet-switching system. To further enhance security, packets transmitted through this network are encrypted to such a degree that only the National Security Administration can

break the code.

New test types make content theft impossible. For example, the item bank of a properly designed computerized adaptive test may contain several thousand items, of which individual examinees are exposed only to a small fraction during each test. Similarly, a live-application test requires an examinee to perform an actual task — a process which is fairly immune to most security concerns. If you can do it, you pass. If you can't do it, you don't!

Q: *How does Web-enhanced testing work?*

A: The host computer (which holds the test items) sends the test to the destination computer (which administers the test) at the testing center and intermediary computers (which transmit the information along the way) communicate with each other through a browser, such as Netscape Navigator or Microsoft Internet Explorer. Browsers interpret universal HTML standards, and so the type of host, destination, and intermediary computer clients, whether Macintosh or PC, are irrelevant. Likewise, operating systems—Windows, MacOS, Unix, OS2/Warp, etc.—are equally unimportant. As a result, any person with an Internet connection can access a test site on the Web.

In traditional Web-enhanced testing, test questions are delivered in real time, subject to the limitations of the Internet's low bandwidth, i.e., the narrowness of the tube that data can flow through. This method of delivery currently can result in delays between items and the slow appearance of graphics.

In this regard the CATGlobal(tm) Testing Network takes a different approach to using the Web for test delivery. In this case complete tests are sent electronically to a proctored test center (or anywhere a test is needed). The test is taken locally and does not require continuous Web access. Therefore it is not subject to the unpredictability of the Internet.

Q: *Where can it take us?*

A: With Web-enhanced testing, examinees have the freedom to register for tests online, at any time of the day by logging into a particular Web site. Registration takes minutes, not hours or days. Paperwork will not get lost in the mail and test candidates do not need to register in person at the testing center. Because the Web delivers tests in seconds, tests can be offered daily, registration can be immediate, and recipients and sponsors can receive score reports in real time.

The use of the computer as a mechanism for test delivery allows the testing of virtually any skill through myriad modalities, including live-application and simulation-type tests, and computerized adaptive testing (CAT).

The greatest advantage of any type of Web-enhanced testing scenario is the ability to deliver the latest tests and return the results in the least amount of time at minimal cost. Should a test developer change a test in the morning, it can be updated for the very next test taker. And when the test is complete, results are returned to the sponsor immediately.

Richard C. Gershon is President and CEO of Computer Adaptive Technologies, Inc., a leading provider of computer-based solutions to testing and survey organizations worldwide.

Special thanks to Laini Wolman, Technical Editor at CAT, Inc., who contributed to this article.

Richard C. Gershon, Ph.D.

President, CEO, and founder of CAT, Inc.

Dr. Gershon has been a leader in testing and testing automation for over 16 years. He has published articles and presented papers nationally and internationally on a wide range of testing issues.

He has served as a program chair and discussant for conferences across the globe. Dr. Gershon holds numerous copyrights for software algorithms used in the testing automation process. He is the former Director of the Northwestern University Testing Center and continues to serve as an adjunct faculty member.

"What the human sciences require for more dramatic progress [is] not simply more data (of the same kind), as so many empiricists have stated, but new instrumentation for obtaining data, or reasonable theoretical restrictions of data domain so that more exhaustive explanatory possibilities can be tried." Ackermann, John R. Data, instruments, and theory: a dialectical approach to understanding science. Princeton, New Jersey: Princeton University Press, 1985, p. 169.

How Good Was Bobby Fischer In 1992?

John Michael Linacre, Ph.D.

In 1992, former world chess champions Bobby Fischer and Boris Spassky met in a grudge match to finally answer the question, "Who is the better player?" They previously met in single combat. Fischer defeated Spassky to become world chess champion. But the match was more an exhibition of gamesmanship than chess. Fischer retired from competitive chess shortly afterwards.

In 1992, they met again. Once more Fischer prevailed with 10 wins, 5 losses, and 15 draws in 30 chess games. Fischer was no longer internationally ranked, but Spassky's proficiency was deemed to be 2560 international master points. World champion Gary Kasparov was then rated at 2780 master points. How would Fischer have fared against Kasparov?

The proficiency of the leading international players in 1992 can be ascertained from their performance in the top ten tournaments reported in Chess Informant. In most tournaments, 10 or 12 players participated, each playing all others present. The result of each match was recorded, as well as the international ELO points standing of each player according to their career performance. In these tournaments there were 88 different players, with 40 different international standings. Neither Fischer, Spassky, nor Kasparov participated.

The outcome of each encounter between two players was entered as a paired comparison into a data file for analysis by the Facets computer program. The players were identified by their international standings, so that the measure corresponding to each international standing could be estimated from a many-facet Rasch analysis of player performance. Results are shown in Figure 1. Each "X" in the figure corresponds to one of the 88 tournament players. The most proficient player in the tournaments was Anatoli Karpov with a standing of 2715. His measure was 5.4 logits, relative to the overall mean performance level of the 88 players which was set at 3.0 logits.

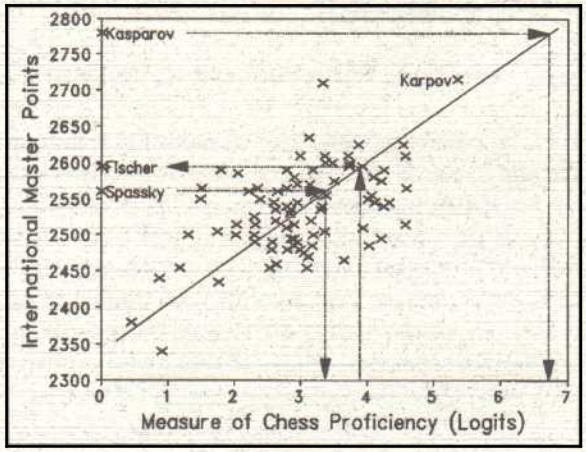


Figure 1. Measures and international standings of 88 leading chess players.

The diagonal line in Figure 1 is the best-fit line between the international standings and the logit measures. Its slope is 66 international standing points per logit. From this plot, Spassky's 2650 standing would give him an expected measure of 3.4 logits. Kasparov's 2780 corresponds to 6.7 logits. Figure 2 plots match outcomes conceptualized as rating scale categories (win, draw, loss), based on the performances of the 88 tournament players. Fischer's raw score of 171/2 against Spassky's 121/2 places him as .5 logits more proficient than Spassky, i.e., at 3.9 logits. Thus Fischer's estimated international standing is 34 better at 2594 points.

Figure 2 would predict Fischer's outcome in a 30-game match to be 8 wins, 19 draws and 3 losses = 171/2. This is reassuringly close to the observed outcome of 10 wins, 15 draws, and 5 losses. This relative lack of draws by Spassky and Fischer may be explained by players being more ready to agree to draws in tournaments when the overall winner is no longer in doubt.

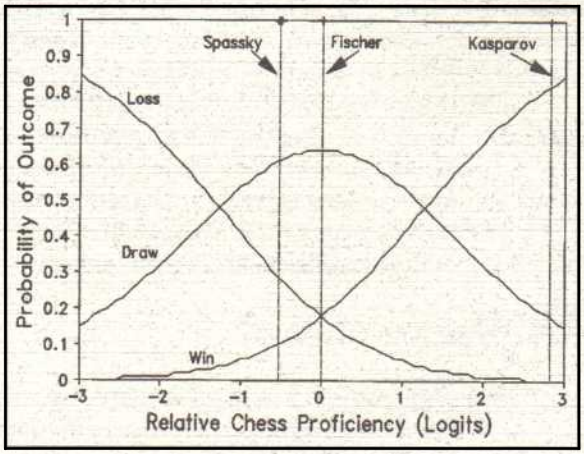
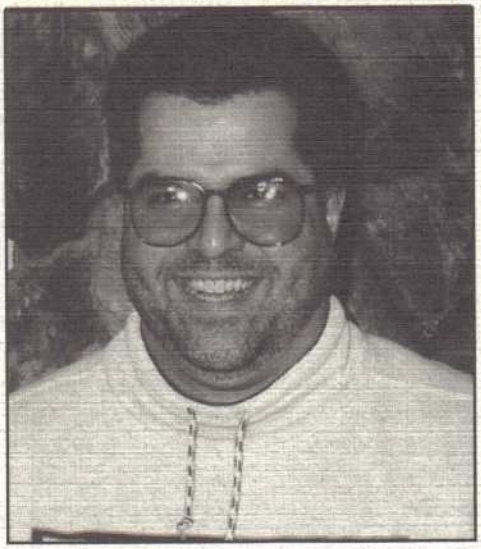


Figure 2. Probability of match outcomes.

World champion Kasparov had a standing of 2780. This is 2.8 logits above Fischer's estimated 2615. Fischer's performance against Kasparov can be predicted from Figure 2. The results of a 30-game match would be 5 draws and 25 losses for Fischer without any wins! Nevertheless, Fischer could be proud. Despite his almost 20-year absence from tournament play, Figure 1 shows that his performance would place him among the top twenty of these tournament players.

Objective Analysis Of Golf

Patrick Fisher, M.A.



With the emphasis on who is truly the best increasingly debated, outcomes measurement has finally made its way to sports performance. Many potential applications of outcomes analysis are available: baseball players, college sports polls, competitive figure skating, and almost anything related to sports that currently is evaluated. Some of the more complicated problems may take years of research to arrive at a complete answer, while others, much less difficult, can be analyzed quite simply.

Of all sports measurement problems, those presented by the game of golf are probably the easiest to solve due to its scoring method. This FACETS analysis is of the hole-by-hole scoring of the 1990 United States Open at Medinah Country Club, Medinah, IL in August, as reported by the United States Golf Association (USGA). These data were collected over the four-day tournament as the players turned in their score cards.

Table 1 shows the players in order of ability in this particular championship. The winner, Hale Irwin, is at the top, but he finished regulation play in a tie with Mike Donald. Irwin won in a subsequent sudden-death playoff, after finishing in another tie following an 18-hole playoff round.

Table 1 - Persons Measurement Report

Measure	Error	Persons	
0.46	0.20	Hale Irwin	BEST
0.46	0.20	Mike Donald	
0.38	0.20	Nick Faldo	
0.38	0.20	Billy Ray Brown	
0.34	0.20	Mark Brooks	
0.30	0.20	Greg Norman	
0.30	0.20	Tim Simpson	
0.30	0.20	Steve Jones	
0.30	0.20	Scott Hoch	
0.26	0.20	Craig Stadler	
0.26	0.20	Tom Sieckmann	
0.26	0.20	Jose M. Olazabal	
0.26	0.20	Fuzzy Zoeller	
0.26	0.20	John Inman	
-0.10	0.19	Tom Kite	
-0.10	0.19	Blaine McCallister	
-0.10	0.19	David Duval	
-0.13	0.19	Bob Gilder	
-0.16	0.19	Scott Verplank	
-0.19	0.19	Ronan Rafferty	
-0.23	0.19	Robert Gamez	
-0.26	0.19	David Graham	
-0.29	0.19	Howard Twitty	
-0.33	0.19	Brad Faxon	
-0.53	0.18	Michael E. Smith	
-0.59	0.18	Randy Wylie	WORST

Table 2 shows the days in order of difficulty to achieve a good score from the hardest, Sunday, to the easiest, Friday. In theory, the difficulty order of the days would be Sunday, then Saturday, Friday, and Thursday as the easiest. Sunday should be the most difficult day because psychological pressure is most intense on the final day of scoring, when tournament ends and the championship is decided. This analysis shows that theory to be essentially correct. Thursday and Friday were misordered, but only slightly, as their measures were only .03 apart. As expected, this analysis shows Sunday the most difficult day by a significant margin.

Table 2 - Day Measurement Report

Measure	Error	DAY	
0.28	0.05	Sunday	HARDEST
-0.01	0.05	Saturday	
-0.12	0.05	Thursday	
-0.15	0.05	Friday	EASIEST

Reliability 0.92

Table 3 shows the holes in measure order from the hardest hole on which to achieve a low (good) score to the easiest. Holes 12 and 16 were hardest to get scores under par, and Holes 14 and 5 were easiest on which to score well. Reliability is very good for the holes calibrations (bottom of Table 3, .92). This table provides useful data for golf course operators wanting to handicap this course fairly for non-championship use.

Table 3 - Holes Measurement Report

Measure	Error	Holes	
0.56	0.09	Hole 16	HARDEST
0.49	0.10	Hole 12	
0.32	0.10	Hole 18	
0.29	0.10	Hole 6	
0.28	0.10	Hole 9	
0.27	0.10	Hole 4	
0.22	0.10	Hole 17	
0.12	0.10	Hole 15	
0.02	0.10	Hole 2	
0.00	0.10	Hole 8	
-0.04	0.10	Hole 3	
-0.04	0.10	Hole 13	
-0.13	0.10	Hole 1	
-0.15	0.10	Hole 7	
-0.34	0.10	Hole 10	
-0.37	0.10	Hole 11	
-0.70	0.10	Hole 14	
-0.80	0.10	Hole 5	EASIEST

Reliability 0.92

In Table 4, the bolded portion demonstrates the effect performance pressure had on two players. Brad Faxon and Ian Woosnam both shot the same score on the same hole, but on different days. Faxon shot a 3 over par 6 on Sunday, the most difficult day, while Woosnam shot the same on Friday, one of the two easiest days. However, the table shows Faxon with a standardized residual of three and Woosnam with a five. Thus, Woosnam's performance was more unexpected, more of a surprise than was Faxon's. There are two reasons for this difference. First, Faxon placed second from last (13-over par); so a bad score would have been more expected from him than from Woosnam. Second, Faxon shot this on Sunday, the day bad scores were expected more frequently than any other day.

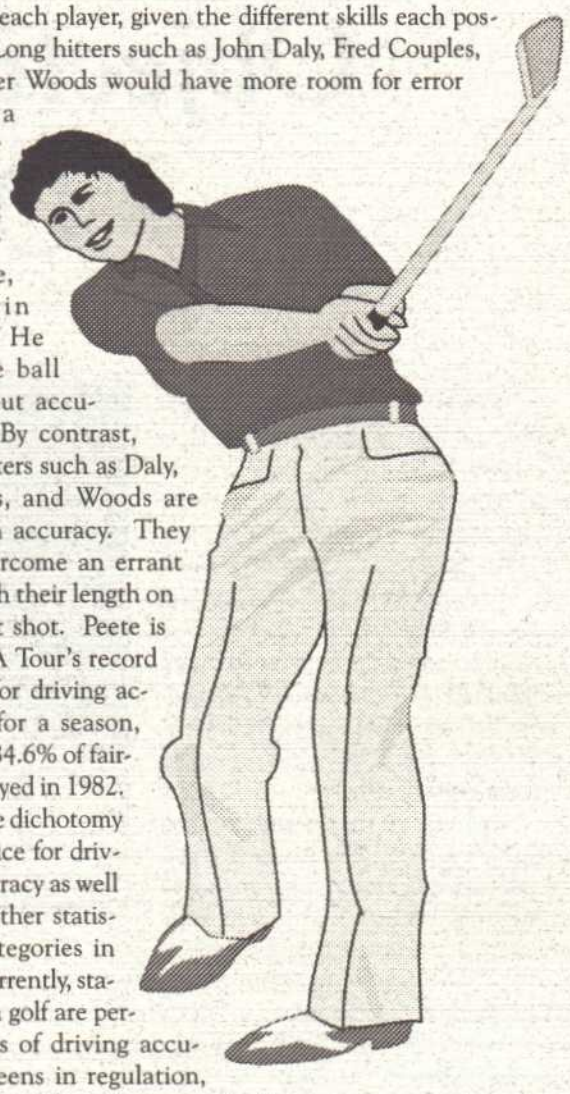
Table 4 - Misfitting ratings

StRes	DAY	Persons	Holes
3	Thursday	Jose Maria Olazabal	Hole 17
3	Saturday	John Huston	Hole 17
4	Saturday	Scott Simpson	Hole 17
5	**Friday**	**Ian Woosnam**	**Hole 17**
3	Saturday	Ian Woosnam	Hole 17
2	Sunday	Chip Beck	Hole 17
2	Sunday	Andy North	Hole 17
2	Sunday	Lanny Wadkins	Hole 17
3	**Sunday**	**Brad Faxon**	**Hole 17**

On each of the four tournament days, the pin placement is changed on each green. This is to prevent the players from becoming too familiar with each hole and increasing their knowledge of how best to play the hole. It is done at the discretion of tournament officials; however, there are no daily increments to make one day harder than another. In a pre-Open article in "Golf Magazine" (Golf, June 1990, pp. 114-124), Curtis Strange, two-time defending champion of the U.S. Open, identified five holes which "will play a part in deciding who wins the Open." From this statement we may surmise that these are the most difficult holes in the tournament. He chose Holes 4, 7, 12, 13, and 16. On the FACETS analysis, Holes 12 and 16 came up to be the most difficult. Thus, Strange had predicted only two out of the top five "hardest" holes to play.

However, when looking at actual scores, Strange's forecast was correct to some extent. The second and third place finishers, Mike Donald and Nick Faldo, respectively, both shot a bogey on Hole 16 on Sunday that would have given Donald the championship and Faldo would have qualified for the playoff with Donald and Irwin. On the other hand, tournament champion Hale Irwin parred Holes 4 and 16 and scored birdies the other three holes on Sunday. He shot 5-under for the day, which set him up for the opportunity to win the playoff. Five-under par was the second lowest score over the four days. Thus, Strange was partially correct about his selected group of five holes that would "play a part" in the decision of the winner.

This analysis is simple, but a more detailed analysis is possible. Each golf stroke results in a task done correctly or incorrectly, (e.g., in the fairway or not). Certainly there are varying degrees of "correctness" — but those that digress also vary for each player, given the different skills each possesses. Long hitters such as John Daly, Fred Couples, and Tiger Woods would have more room for error than a player with the different skill, for instance, ofCalvin Peete. He hits the ball short, but accurately. By contrast, long hitters such as Daly, Couples, and Woods are lower in accuracy. They can overcome an errant shot with their length on the next shot. Peete is the PGA Tour's record holder for driving accuracy for a season, hitting 84.6% of fairways played in 1982. A simple dichotomy will suffice for driving accuracy as well as the other statistical categories in golf. Currently, statistics in golf are percentages of driving accuracy, greens in regulation, and saves. These factors and a few more have an impact on the score earned on each hole. These factors in golf could be analyzed to provide a more comprehensive diagnostic view of players' areas of weakness and strength.



This kind of analysis can be helpful to golf course administrators and players. The players could learn more definitively where their weaknesses lie (driving, the short game, putting) and learn how the layout of the course can affect their play. Course officials could be provided with more accurate and detailed data on difficulties of holes existing, or planned for. Such analyses could assist architects in the design of future courses.

Patrick B. Fisher, MA

Mr. Fisher earned his Master's from The University of Chicago in 1993. His field of study was Measurement, Evaluation, and Statistical Analysis focusing on sports performance measurement. His Master's paper was on measuring baseball performance. Mr. Fisher is currently employed by the Rehabilitation Institute of Chicago in the Rehabilitation Services Evaluation Unit as a Program Evaluator & Statistician. He is the proud papa of Bradley Patrick and Brandon Michael born on October 10, 1997. E-mail: p-fisher@nwu.edu.

Assessment:

What is it? Why do we need it? How do we use it?

Assessment is one of those concepts that sounds simple until it is time to design and use an assessment instrument. In order to discuss it, we might ask of the process: What is it? Why do we need it? And how do we use it?

Roy Berko , D.Ed. & Linda Webster, Ph.D.

What Is It?

It is the purpose of the assessment process to develop a tool or measurement device which, when applied, evaluates what we are intending to assess. This circular-sounding description can be reduced to: a test tests what the test intends to test. Or, assessment assesses what the assessment procedure intends to assess. Therein lies the problem with the assessment process. Many schools, departments, and instructors don't know what they want to assess.

A survey by Ellen Hay, reported in "A National Survey of Assessment Trends in Communication Departments," July, 1992, Communication Education, indicated that only a third of these departments defined goals and objectives for themselves. This means they have no clear goal attainment to assess. In addition, many instructors develop courses with no clear specific learning and expectancy goals. Many of those same instructors lack any test and measurement courses or experiences, and so do not have the slightest idea of how to develop assessment tools.

So, we have a problem. Many of our colleagues start with an unclear purpose and then find themselves unable to work toward accomplishing that unclear purpose. Even when they have a clear purpose and the will to accomplish it, they may not know how to set up a procedure to assess that purpose.

In our field, we are expected to add the burden of evaluating skills and concepts which, in many instances, we cannot prove work. In public communication, for example, why is it that student evaluation of "the best" or an "A" speech often does not correlate with ours? Why is there no absolute winner in speech contests? And why couldn't Bob Dole's speech advisors for the 1996 Presidential campaign "make" his speeches work?

Group discussion is another example. How is it that a group refusing to follow an agenda we have made them develop is still able to complete the task? And, finally, we need to consider the ethical dimensions in the evaluation of communication. Can we accurately evaluate human acts? Perhaps it is worth considering that the human tendency toward subjectivity rather than objectivity might get in the way of evaluating communication behaviors. Even more profound, how does one determine the benchmarks for the evaluation? Do we use grading forms that may judge the skills that students brought with them rather than those skills learned in class?

Two students in gym class are required to shoot seven out of ten baskets to pass the class. One student has played basketball for many years and consistently "hits" seven or more baskets from the first day of class. The other student has never played the game and shoots only one or two baskets on an infrequent basis at the beginning of the basketball unit. But this student became more consistent and accurate by the time the coach was ready to grade their performance. The more proficient young man hit his usual seven baskets and earned his passing grade. The less proficient young man made five of his ten baskets and failed the class. Now, if you were grading on improvement or mastery based on what was taught, how would you rate the second young man?

Can grading forms used this way be an accurate tool? What will it take to come up with inter-rater reliability? Are the questions on the grading form the essence of the real display of effectiveness of learning?

Why Do We Need It?

One of the obvious reasons for needing assessment is that teachers have to give grades. Coupled with the semester-end assessment in the classroom is the pressure for performance testing at all academic levels from state legislatures and Departments of Education. Many institutions are moving toward individual exit competencies for their majors including capstone courses, testing, and portfolios.

Additional pressure comes for outcomes-oriented teaching assessment at the collegiate level brought by accreditation agencies. For example, southern collegiate institutions must graduate communicatively competent students, though no

definition is included as to what that means.

Beyond assessing the individual student is the move toward assessing whether our departments, schools, and programs are fulfilling their missions, a particularly tough assignment for those schools without mission statements. Then there are the "housekeeping" roles of assessment, such as proficiency testing for waiver credit and placement testing for communication courses.

How Do We Use It?

Our greatest need is to prove that our courses are accomplishing their objectives. The Hay study, "A National Survey of Assessment Trends in Communication Departments," indicated that 66% of the institutions in the survey included "communication skills" in their general education requirements, and assessment was used to prove that learning had taken place. How? 83% indicated that by passing the communication requirement, a course or courses, the students had proven that they were competent. The other 17% required their students to pass a specific performance or test.

Some schools like Radford and Hamline University are more specific, requiring that students demonstrate their communication proficiency in a variety of contexts over an extended period of time. Other institutions, such as Golden West College, go further by having laboratories where students are required to prove their skills and knowledge through a series of performance activities.

We also need to prove to accrediting agencies that the school/program is reaching its required goals and to certify that their majors have learned the necessary materials and have developed the required skills in the completed courses.
The Hay study also indicated that constituents from other fields have an interest in the development of oral communication assessment. It was found that 49% of the states require teacher education programs to include an oral communication component. It is interesting to note that one of the highest levels of communication apprehension within occupational groups is that found among elementary teachers, the very people we expect to teach communication skills to young children. Additionally, organization such as ASTD (Association for Training and Development) is looking to our field of communication for teaching and assessment models.

We need to work on answers to these questions. While this is only one side of the dialogue both within, and without, the field of oral communication, it is a dialogue that is both timely and pressing.

The work done by Donna Surges Tatum and her colleagues at the University of Chicago provides many of the answers for our vexing questions. We need to listen with care and implement the scientific principals developed for performance assessment. By doing so we enhance the credibility of Communication Studies as a discipline of both the Arts and Science.

Roy Berko

Roy Berko is a Senior Communication Consultant with Martel and Associates. He was formerly a visiting Professor at George Washington University, an Associate Director with the National Communication Association, and a Professor at Towson State University and Lorain County Community College.

A graduate of Kent State, University of Michigan, and Pennsylvania State University, he is a certified Counselor, hypnotherapist, and negotiator, and has been in private practice as a psychological counselor.

Dr. Berko is the author or coauthor of over twenty books and numerous scholarly articles. He is a nationally recognized expert in the field of communication who has appeared on such programs as *Good Morning America* and *Fox Morning News*, and for three years served as the communication expert for ABC-TV in Cleveland, Ohio. He has also appeared regularly on National Public Radio and served as a Public Relations Advisor to the Volunteer Office at the White House.

He has received five national teaching awards, including the prestigious Teacher on Teaching from the National Communication Association and Master Teacher Recognition from the National Conference on College Teaching and Learning.

There is nothing more difficult to plan, more doubtful of success, nor more dangerous to manage than the creation of a new system. For the initiator has the enmity of all who would profit by the preservation of the old institutions, and merely lukewarm defenders in those who should gain by the new ones.

Machiavelli

Public Speaking Assessment for College Students

William W. Neher, Ph.D.
Deborah Grew, M.A.

Meaningful Measurement (MM), a system devised by Donna Surges Tatum based on Communication theory and a mathematical model, produces objective measures of student performances. This technique allows us to compare evaluations across sections and courses. We should thus be able to document real improvement in competence for individual students as well as for groups of students, regardless of the persons doing the rating. The method can provide evidence for actual "value added" for a given assignment, course, program, or curriculum when used cumulatively (Tatum, 1997).

Assessment through MM has come to our university at a propitious time. The university is embarking on a major initiative on student learning outcomes, and the implementation of MM has been funded by the Lilly Foundation. Our "learning initiative" is intended to direct attention to measuring student progress in terms of outcomes, what they actually know and can do, rather than in terms of hours or courses completed (the "inputs" approach to charting student progress). The Lilly Foundation has provided grants for several private colleges and universities to enhance the effectiveness of the transition from high school to post-secondary education. Butler's grant is divided among several initiatives, two of which are Communication-Across-the-Curriculum and Meaningful Measurement.

The results of our pilot study here based on an analysis of the use of MM in eight sections of basic public speaking indicates that the rating items were reliable and that raters were consistent in their use of the items. Of most interest is that the analysis documents that student speakers exhibit real improvement (well beyond chance) as a result of the courses. The analysis also provides breakdown for improvement from first to second speech, from second to third, and, when possible, from third to fourth speeches in a semester. This issue is of special interest in our department as we are concerned to determine whether there are an optimum number of graded speaking assignments that should be required in a basic semester course. The analysis also provides data indicating the learning outcomes, or assessment, of the course.

During summer 1997, the Communication Studies Department held a workshop concerned with faculty and course development. We took up the matter of expanding the implementation of MM to all sections of SH101. Donna Surges Tatum attended two days of the workshop to help faculty further understand MM. Several important steps were taken at the workshop to broaden the program at Butler University.

First, the Communication Studies faculty discussed the rating form and decided to make some changes with regard to the items used on the form. Changes were made to reflect a more universal consensus of what expectation we have of skills students should master in a public speaking course. Two forms were developed: one with the ratings (1-6) Terrible, Poor, Average, Good, Very Good, Excellent to the right of each item; another was developed for faculty use with a line to write the numbers 1-6 and also a comment area to the right of each item. Second, we rated and discussed videotaped speeches of Butler students in order to examine our rater behavior and to determine what we look for as instructors. Third, we formed a small group of three faculty members to view videotaped speeches from Butler in order to create new norming tapes for use at Butler. Four videotaped speeches were selected to become norming speeches. These speeches were chosen on the basis of completeness, relevance and variety, clarity of speech, and tape quality. The faculty members also looked at delivery, clarity of content, and variety of speaker organizational methods.

All four speeches were delivered as part of a competition we call Speech Night. The speakers competing in the preliminary rounds were voted on by their classmates in each section and were often the better speakers in the class. All speeches were persuasive. The four speeches selected by the faculty panel were then copied onto videotapes for use for norming purposes. Also during summer 1997, a faculty development workshop was offered to faculty outside the Communication Studies Department. Faculty members attended this workshop from the School of Pharmacy, Fine Arts, Business Administration, and the Liberal Arts College. MM was of special interest to pharmacy faculty members because of a course offered in the School of Pharmacy called "Professional Communications" which is designed to help student-pharmacists develop their speaking and consulting skills when discussing medications with patients and their family members.

In consultation with the pharmacy faculty, the MM rating form developed at Butler was modified to be applicable to

their needs. The student-pharmacists were observed and rated using an interview-style form. Items from the MM form were chosen which were most applicable. Nineteen items were pulled out of the SH101 form and the descriptors were changed to focus the items on the needs of the consultation setting.

The "Student-Pharmacist Consultation" form is now being used in both sections of the Professional Communication course. Forty-seven students and five faculty members were normed using four videotaped student-pharmacist consultations, establishing a baseline for the raters (student-pharmacists and faculty) with these individuals becoming connected to the larger database through the same MM items as appear on the SH101 form.

There are four rounds of student-pharmacist consultations during the semester. In each of the rounds, students rehearsed interpersonal skills with different "patients." In Round One, students act as "patients," and students and School of Pharmacy faculty rate the student-pharmacists. In Round Two, other Butler University faculty members and residents of a local retirement community act as "patients." During Round Three, faculty was used as "patients," and the consultations, which are rated by the pharmacy students, are also videotaped, because the student-pharmacists have the opportunity to compete in a national competition. Round Four consists of "live" consultations with faculty members as "patients." Service-learning students, who are students training "in the field" at pharmacies in Central Indiana, also act as consultants and as raters.

The logistics of implementing MM are quite simple. Students are hired for data entry and have responsibility for particular classes. Each faculty member organizes his/her semester differently, so weekly data entry duties are a bit unpredictable, but an average of about fifteen hours a week is spent entering the speech ratings for all twenty SH101 sections and the pharmacy course.

All faculty members have elected to use MM in some manner in their class. Some have every student rate every speech; others have students rotate as raters. Data is e-mailed twice a week to Donna Surges Tatum, and reports are sent back the following day. Each report consists of Overall Speech Measure, and the subscales of Speaker, Audience, and Message measures. Instructions are included to help faculty interpret the report and give useful feedback to the students.

Halfway through the MM project, some observations are possible. Assessment is a faculty development tool. When we as teachers must think about what is being assessed, it forces us to re-examine our teaching, and refine the classroom experience.

The speech measures have a high correlation with the speech grades as given by faculty. Thus the objective measurement is supported by the subjective evaluation. This is of great importance to the skeptics who did not believe that it is possible to produce calibrations and measures in a performance situation such as public speaking. They now see objective measurement as a teaching tool and are willing to participate.

Butler University's commitment to the learning initiative is enhanced when we have a definitive method of assessment. We can pinpoint just how much value has been added to each student who takes this required Public Speaking course.

William W. Neher

Education: Ph. D., Northwestern University, 1970. Communication Studies, Program of African Studies. Dissertation: Public Address in Kenya: A Study in Comparative Rhetoric, Intersocietal Studies grant, research in Kenya, 1969-70. M. A., Northwestern University, 1967, Communication Studies. B.A., Butler University, 1966, History.

Bill Neher is professor of communication studies at Butler University. He has been at Butler for 27 years, where he has served as Dean of the University College, Director of the Honors Program as well as Head of the Department of Speech Communication, now Communication Studies. He is currently the chairman of the Faculty Assembly, the faculty governance body at the university.

He is the author of several books dealing with speech communication and business and professional communication. His latest book is on organizational communication, published by Allyn & Bacon of Boston, The Challenges of Change, Diversity, and Continuity: Dimensions of Organizational Communication. Other works include The Business and Professional Communicator, with David H. Waite, published by Allyn and Bacon in 1993.

In addition to his duties in the Department of Communication Studies, he also teaches in the Butler Change and Tradition core program, the MBA program (courses in organizational communication), as well as courses in African studies. He has served as a consultant and trainer in presentational speaking for, among others, AT&T, PSI Energy, Indianapolis Power and Light Co., the City of Indianapolis and State of Indiana, TransUnion Corporation, Department of Public Instruction, several health organizations, charitable organizations, and professional associations.

Deborah Jean Grew
Director, Computerized Public Speaking Assessment
Butler University

B.A., Indiana University
M.A., University of Montana

Debbi is married with one child and one dog. She enjoys running and exercise and will run in the Indianapolis 500 Mini-marathon for the sixth time this May.

Her favorite travel spots are Maine and Cape Cod.

A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.

Max Plank

Student Progress? Prove It!

Donna Surges Tatum, Ph.D.

Course Goals

Many business and professional people recognize the importance of being able to communicate publicly, because they seek training to improve their skills. Effective communication skills are a highly desired commodity in today's job market. Corporations value such things as team-building, accountability, customer service, total quality management, and 360-degree employee evaluations. That, and the increasingly rapid changes in the workplace, make management acutely aware of the importance of competent communicators. The seas of change are best navigated by those who know how to ask for and give directions.

Butler University responds to this need by offering Public Speaking courses. The purpose of this assessment project is to determine the efficacy of the training Butler provides its students. Careful research design and precise measurement provide the basis for this report.

Demonstrable results in the following areas are the teaching goals of the course:

To enhance delivery skills

To teach methods of organization and critical thinking skills

To increase confidence.

Research Questions

1. Is the evaluation form valid and reliable?
2. Are student raters reliable and consistent when rating their peers?
3. Do students improve their public speaking skills when they take Public Speaking classes?
4. Is inconsistency as a rater related to that person's public speaking ability?
5. Is rater severity related to public speaking ability?

Data Description

The data were collected Spring semester of 1997, from a variety of classes taught by four instructors. One hundred forty-eight students gave 381 speeches which were evaluated by 151 raters using a 29-item, six-point scale instrument. A total of 4925 rating forms are in the database.

Assessment Issues

The assessment of oral communication skills has long been fraught with problems other areas such as math and English do not have. One can administer a test in arithmetic, count the correct answers, compare standardized scores, and come up with a reasonable estimate of a student's ability. The expectations for ability are grade- and age-related, and a common frame of reference has been established over the years.

The communication field is now developing such a clear-cut method of evaluation. This assessment project is using the Meaningful Measurement system which uses the Linacre FACETS extension of the Rasch model as the basis for calculations. It is a method which takes subjective, qualitative observations, and transforms them into objective, quantitative measures. The Meaningful Measurement system is designed to maximize the science of assessment. All raters evaluate four videotaped speeches. This provides common ratings to link and calibrate the raters at this school and others across the country. The rating items are checked for fit and calibration.

The following questions are the psychometric and fairness issues of any situation where raters assess skills.

1. What are appropriate expectations?

What proficiency should be required of a ninth grader, a community college student, or a graduating college senior? Do we know the hierarchy of skills? Have we calibrated the competencies? Do we know which skills should be accomplished at what level and in what order? Our intuition and experience must be backed up with the facts of measurement. The Meaningful Measurement system gives this information to the faculty of Butler University so they can make the proper pedagogical decisions.

2. Are the evaluation instruments sound?

Do the items cover the range of the variable? That is, are there some items that are easier than others? It is not useful if items are bunched together. That would be like giving a test of only simple addition problems. We would not find out the student's true ability, only whether he or she can add. If there is a range of easier to harder items, we can pinpoint with greater accuracy the level of a student's competency.

Do all of the items "fit"? Do they measure what they are intended to? Which items need to be rewritten or dropped? Checking for fit also allows us to be sure we are only measuring one thing at a time, and not confusing issues. (For instance, a story problem on a math test may be more of a reading than math question.) If we are not careful, and try to compare apples to oranges, what we end up with is fruit salad.

The rating form used for this assessment project passed all tests with flying colors. It has 29 items targeted to essential competencies and covers a range of about 90 measure units. The two misfitting items are visual aid quality and use. This is due to the visibility in the classroom, which depends on where the rater is sitting.

3. How are differences in raters accommodated? How do we achieve objectivity?

Assessing oral communication skills most often is done by a teacher, or other trained judges, using a rating scale. We know that we all live in our own perceptual world, and attend to different things. Thus, no matter how hard we try for "inter-rater reliability," we will never achieve the ideal of all raters being equal. Instead of a false assumption of sameness, we must address the issue of differences. The most important factor in rating is the consistency with which the judge uses the evaluation form.

When assessing skills, we must be very careful to ensure objectivity in a situation which is subjective by nature. We must have a mechanism to control for levels of severity as well as bias. Meaningful Measurement adjusts for the variations in severity, and flags an inconsistent or biased rater.

4. How can we compare results?

What does a raw score of "65" mean? For example, students are assessed on a 20-item, 4-point rating scale instrument by several different raters. The next year new students are evaluated by some of the old and some new raters. Can we compare the students to each other? One judge is very easy, and gives high ratings. Are those students' raw scores "worth" as much as the raw scores received by students who were rated by a tough judge? How do you come up with a fair ranking? Are the students this year truly better than the ones last year? How do we know for sure?

Meaningful Measurement calibrates all speakers on the same "ruler." This makes it possible to directly compare students from speech to speech, class to class, or year to year.

5. How does a teacher maintain a stable frame of reference throughout the course?

It is difficult to think back to the beginning of the semester, and pull up an accurate recollection of a student's performance. We usually have a general impression, and perhaps a remembrance of a specific skill or two. Referring back to rating forms may help, but it is tedious and fuzzy.

With Meaningful Measurement a teacher can refer to calibrated measures and know precisely how much improvement has (or hasn't) taken place over the semester.

Results

Units of Measure

When reading Meaningful Measurement reports, all numbers are directly comparable. For example, money is in common units; we all know there are 100 pennies in a dollar and that a "dollar" is a "dollar." A dollar is comparable from year to year. We have a common frame of reference. When Dad reminisces about paying 17 cents for a gallon of gas thirty years ago, we know we're paying about ten times that amount today. We can adjust for inflation to determine what the real differences are, yet still be in the same units of measure. When we go to the grocery store to buy food, then to a restaurant for a meal, the bills are both in dollar units. We can compare the price of the ingredients in a tossed salad with what it costs to buy one at a fancy cafe. Even though the situations are different, we can maintain a common frame of reference for the relative costs.

The same situation applies to assessment. When our reports are given, they are in units of measure called "logits." Each logit can have 100 points and has the same properties as a dollar. We can compare one "logit/price" to another. We can add and subtract with logits. Student A's first speech measure is 10.05, and her second measure is 11.45. We know she has progressed by 1.40 logits, or 140 points.

The scale has been calibrated so the origin, or balance point, is "10.00." That means a speech which is of average ability, or a rater who is of average severity, has a measure of 10.00. The lower the number, the less able or less severe a person is measured. Measures higher than 10.00 indicate more ability or severity than that of the "average" speaker or rater.

We have established and maintained a metric that can be used from year to year, and situation to situation. We have the means to track and assess improvement.

Raters

The 151 raters are examined to determine how consistent they are when rating speeches. An investigation of the fit statistics shows that 84% of all raters are "good." That is, they are internally consistent and are able to maintain a stable frame of reference when evaluating speakers. This means we can trust the speech measures. The raters are not behaving erratically.

The raters' mean severity measure is 10.00. They fit well, but cover a wide range of severity from easy to hard when rating speeches.

Items

The Item Map below shows the hierarchy of items. The Butler University speech communication faculty determined that these are the essential competencies required of the students when giving a speech.

The calibration of the items goes from easy to hard. The lower the number, the easier the item is to accomplish. The items cover a range of 95 points. The point biserials show that all the items are related, and define a common variable. The separation reliability is .99.

At Level 1 the easist thing for the students to do is to show their knowledge/mastery of the topic, pick a worthy topic, and appear trustworthy.

At Level 2 the next easiest items include showing the relevance of the topic, using appropriate language, being understandable, using materials appropriate to the audience, limiting the topic, and using clear language.

At Level 3 the visual impression of the speaker, word

choice and establishing common ground are a bit more difficult. A well-organized speech using good quality support are next in the hierarchy.

At Level 4 ethical and appropriate emotion appeals are slightly above average in difficulty, as are eye contact and a poised demeanor.

At Level 5 a conversational style and variety in vocal delivery are more difficult to accomplish. The quality and use of visual aids are also in this strata.

It is progressively more difficult to use a sufficient quantity of verbal support with a variety of sources, and to respond to audience feedback. Well-presented support with citations and establishing a context is harder to do.

At Level 6 an enthusiastic delivery is quite difficult on this scale. The flow of the speech with preview/review, signposting, and transitions is also at this point.

Finally, **at Level 7** fluency and smoothness in vocal delivery is the second most difficult thing for a speaker to do. Gestures are the hardest for a speaker to effectively accomplish at Level 8.

Speech Results

Ninety-four students in the basic course gave at least two prepared presentations, 88 gave three, and 11 gave four. Thirty-two students in the advanced course gave two prepared presentations.

The mean of all speeches is 11.64, or 164 points above the mythical average speaker at 10.00. This shows the Butler University student body is an accomplished group. The separation of 8.18 and standard deviation of .75 demonstrate there is a wide range of ability in this sample. The normal, bell-shaped distribution shows speakers' ability from about 8.20 to 13.60, a range of over 500 points.

ITEM MAP

EASY	SPEAKER	MESSAGE	AUDIENCE
1	mastery trustworthy	worthy topic	
2	understandable	appropriate language limit topic clear language	relevance materials appropriate
3	visual impression word choice	well-organized	common ground
4	eye contact demeanor		ethical emotion appropriate emotion
5	conversational variety	aid quality aid use quantity support	responds to feedback
6	enthusiastic	well-presented support flow of speech	
7	fluency		
8	gestures		
HARD			

Speaker Improvement - 2 Speeches

Ninety-four students gave two prepared presentations. The mean measure for the first speech is 11.17. The second speech measure averages 11.45. This is an average gain of over a quarter of a logit, or 28 points.

A paired samples t-test tests the hypothesis of whether the first round of speeches is the same as the second round of speeches.

In other words, does training make a difference? Do speakers improve? The answer is "Yes!"

The t-value of 4.56 with a significance of .000 means we are absolutely sure: The two groups are truly different, and the improvement is not due to chance.

Speaker Improvement - 3 Speeches

We know students significantly improve from their first to their second speeches. Now we want to know if they continue to gain in ability.

Learning does not stop after two rounds of speeches. Students have not learned all there is to know about public speaking after just two speeches, for they continue to improve as shown by the following table.

Seventy-seven students gave three prepared presentations. The results of this group are shown, for instance, through the

paired samples t-test of the second and third round of speeches.

The mean of this group of second speeches is 11.49, and the mean of the third is 11.71. Again the students improved — this time by .22 logits, or 22 points.

The significance of .000 means we are 100% sure the third round of speeches is truly different from the second round.

Speaker Improvement - 4 Speeches

Eleven students gave a fourth speech. These students improved another 30 points. The t-value of 2.33 with a significance of .045 means we are 95.5% sure that the fourth round gain is due to training.

Speaker Improvement - Advanced Class

Thirty-two students in the advanced classes gave two prepared presentations. These students continue to improve by 35 points. (In reality this is the fourth and fifth speeches for these students because they already had the basic course.) The t-value of 4.08 with a significance of .000 means we are absolutely sure the advanced training has an effect.

Rater Consistency and Speaker Ability

A Mean square (MNSQ) fit statistic evaluates the consistency of the rater. A mean square of 1.0 is exactly what is expected; .7 to 1.3 is normal. But a mean square of 1.5 means there is 50% more "noise" in a rater's evaluations, and 1.9 90% more variance than expected.

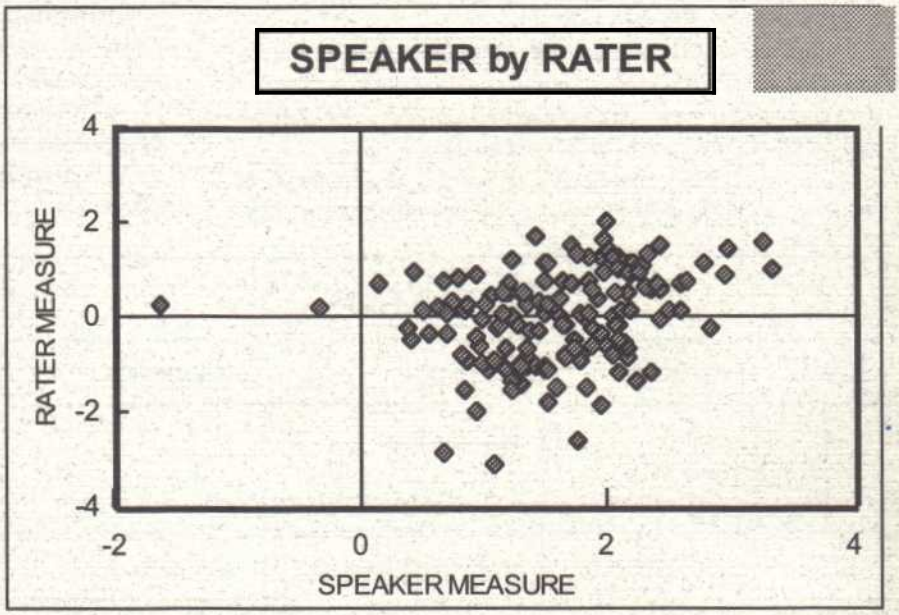
A rule of thumb is to look closely at any response pattern with a mean square of more than 1.4, or a standardized fit over 2. When this occurs, a red flag waves in the researcher's mind, and a close examination of the data is warranted to determine the cause of the misfit. It may be that the rater is consistently inconsistent and should not be used for assessment purposes, or perhaps the rater had a bad day.

Some raters have mean squares and fits that are almost too quiet, mean squares of .5 or below. They are close to Guttman-like in their consistency. Their evaluations hold no surprises or randomness. They are rating holistically instead of discriminating among the items.

Fifteen of the 152 raters are inconsistent, and 10 are overly consistent. The table above shows these 25 rater fit statistics with their speech measures. But there is no relationship between a rater's consistency and speech ability.

Rater Severity and Speaker Ability

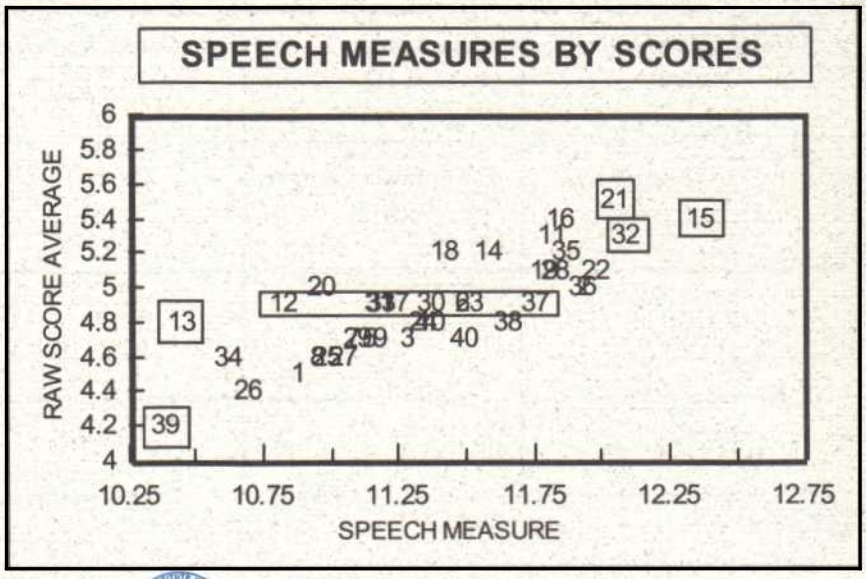
The graph below shows there is not a clear relationship between a person's severity as a rater and their ability as a speaker. Some excellent speakers are easy raters, and some poor speakers are quite severe.



Measures and Raw Scores

The next graph demonstrates the importance of objective measures rather than a proportion of raw scores. When the severity of the rater is taken into consideration, the results can be different.

Forty speeches were randomly chosen from the database. The average of the raw scores is plotted against the speech



measure. Eight speakers have a raw score of 4.9. However, their measures range from 10.82 to 11.75, a difference of 93 points.

The worst speech is #39 with a raw score of 4.2 and a measure of 10.39, yet the second lowest speech, #13, has a measure of 10.45 and a raw score of 4.8.

Speech #21 has the highest raw score, 5.5, but is third in ability after the raw scores are conditioned into measures (behind #32 with 5.3 and 12.09, and #15 at 5.4 and 12.36).

Now we have a method to not only ensure, but prove fairness in the judging process. This is extremely important in grading and other high-stakes assessments.

Discussion

Meaningful Measurement Results

The results show that training in Public Speaking produces positive results. Students significantly improve from their first to second speeches, and they continue to do so in subsequent speeches and in subsequent advanced classes.

We can have confidence that these outcomes are not dependent upon a particular teacher, because the students came from eight classes taught by four different teachers. The Butler University Speech Department is fulfilling its mission, and should be commended for the excellent job it is doing in training its students.

This study also demonstrates:

1. Students are useful, reliable raters. Since audience analysis is taught as an important factor when preparing a speech, we can now derive speech measures from the entire class instead of only one grade from one teacher.

2. Averaging raw scores does not produce reliable speech measures.

3. A student's consistency as a rater is unrelated to his or her ability as a speaker.

4. A student's severity as a rater is unrelated to his or her ability as a speaker.

5. The hierarchy of item difficulty improves our concept of what is required for public speaking ability. Now it is possible to identify the items that turn a poor speaker into a good one. Expectations for progress can be realistic and predictable. Teaching methods improve because information can be sequenced according to actual student development.

Health Care Outcome Measurement

William P. Fisher, Jr. Ph.D.
LSU Medical Center, New Orleans

"The organizations that recognize the challenges, opportunities and rewards of measuring clinical outcomes will emerge as and remain market leaders." from "Clinical Outcomes: The New Driving Force in Health Care" by Raul A. Trillo, MD, Senior Health Care Consultant, Deloitte & Touche Consulting Group, New York, appearing on page 17 of the October 27, 1997 issue of American Medical News.

As everyone is well aware, health care costs are increasing at several times the general rate of inflation. Most health care consumers are also aware that health maintenance organizations (HMOs) are managing care in an effort to slow the spiraling costs, most usually by restricting access to care, as when referrals are required for specialist consultations, or when clinicians are required to follow procedural regimens in the care they provide.

What is less widely understood, however, is that HMOs and managed care produce, on average, only a one-time 7–9% reduction in costs, after which the increases continue unabated. Most approaches to cost reduction taken to date follow the model of quality control, in which the low-quality tail of a quality distribution is lopped off, with no overall change in the structure, process, or outcome of the care provided.

In contrast with the quality control approach is the quality assessment and improvement approach, in which the entire quality distribution is moved toward a higher standard. It is crucial at this point to recognize that costs and outcomes are opposite sides of the same coin. It is impossible to change anything that reduces costs without also affecting outcomes, and vice versa. The point is to be able to evaluate the relation between cost and outcomes in ways that are sensitive to both the organization's mission to provide care and its bottom line.

Outcome measurement systems make it possible to show how much change in health or functioning is obtained per unit cost, and outcome measures have been focused on serving this accountability need, especially in the area of physical medicine and rehabilitation. The key to better outcomes per dollar is process improvement, but it is impossible to evaluate the effect of changes in processes unless outcomes are measured with high reliability and validity.

The vast majority of outcome measurement systems proposed to date mistakenly treat raw, ordinal summed scores as linear, interval measures. Accordingly, the various efforts underway ostensibly aimed at standardizing outcome measures in health care focus on the hopeless task of devising a single collection of items that will meet all users' needs. Though recognition of probabilistic measurement models in research publications is growing (see bibliography), there is not yet much widespread appreciation in health care for the strengths of models that 1) test data quality and the hypothesis that the variable is quantitative; 2) express each facet of the measurement design (item difficulties, person measures, rater harshness/leniency) in a common quality-assessed-and-improved metric; 3) accommodate missing data; 4) facilitate adaptive instrument administration, which adapts technology to the needs of people instead of vice versa; 5) remove from the measures rater and other identifiable and consistent bias factors that can be included in the model; and 6) provide a basis for standard metrics, i.e., universally-recognized, variable-specific quantities that can be read off any calibrated instrument shown to measure that variable.

It is often instructive to observe where things have been if one desires a sense of where they are going. Outcome measurement research in health care employing Rasch's probabilistic models had its first applications in mental health and psychiatry, in the 1970s in Europe and North America (Hehl & Nussel, 1975, 1976; Kalinowski, 1985; Lewine, Fogg, & Meltzer, 1983; Maier & Philipp, 1986; Olsen & Savroe, 1984; Sørenson, Hansen, Andersen, et al., 1989). In the late 1970s or early 1980s, Ross Lambert, MD, an ophthalmologist at the Hines VA Hospital west of Chicago, and Benjamin D. Wright, PhD, became acquainted during early morning swims at a Hyde Park pool.

Lambert was involved in rehabilitating veterans suffering from low vision problems caused by accidents, diabetic retinopathy, or other problems. He needed an assessment tool that would enable therapists to document how well someone with severe visual impairments could perform travel activities, such as walking around at home, in the local neighborhood, in new places, as well as taking a bus or train, using an elevator, or shopping. University of Chicago graduate students, including Larry Ludlow, Matthew Schulz, Sheila Courington, David Zurakowski, Mark Wilson, Patrick Fisher, and this author worked as research assistants at Hines as a result of Lambert's interest in Rasch measurement.

In 1985, Lambert decided to become "double-boarded" and add a professional certification in physical medicine and rehabilitation to his ophthalmology certification. He became part of the first class of residents to rotate through Marianjoy Rehabilitation Hospital & Clinics, also in Chicago's western suburbs. At Marianjoy, Lambert learned that Medical Director, Richard Harvey, MD, had devised a rating-based functional assessment system, the Patient Evaluation Conference System, for monitoring the outcomes of care. Harvey took an immediate interest in testing data from the PECS system to see if they could meet the requirements for measurement specified in a Rasch model. He and Lambert used Wright's software to analyze the data. They presented the results to the Academy of Physical Medicine & Rehabilitation in 1987 (Harvey & Lambert, 1987; Lambert & Harvey, 1987; Lambert & Harvey, 1988; Lambert & Rao, 1989; Lambert & Wright, 1989; Lambert, Yokoo, Kilgore, et al., 1990).

Following the success of these initial analyses, Harvey brought in Burton Silverstein, PhD, in late 1987 to continue the work. Silverstein had just finished a post-doctoral fellowship at the University of Chicago. Harvey and Silverstein saw that the Rasch measurement research agenda held great potential for improving the PECS's capacity to support program evaluation and quality assessment applications, so in April, 1988, Karl Kilgore, PhD, was hired as Director of Research and Education at Marianjoy, and in August this author started as Research Associate. In 1989, Silverstein, Kilgore, and Fisher published a monograph on patient tracking and outcome assessment (Silverstein, Kilgore, & Fisher, 1989). Over the next several years, they together and separately published several articles on functional assessment in rehabilitation, and made many presentations on the topic.

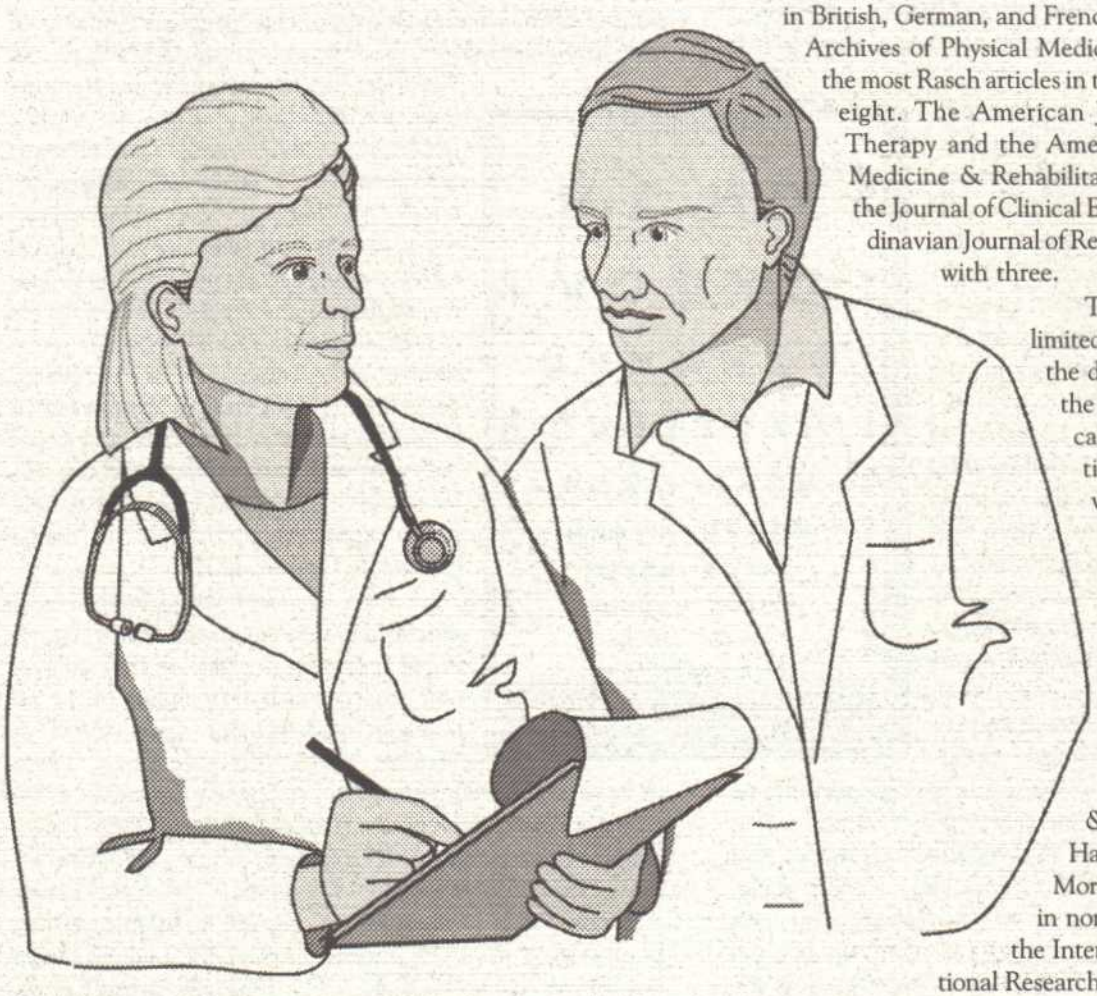
With Harvey as editor and the submission of articles reporting advanced measurement research employing functional assessment instruments, the Archives of Physical Medicine and Rehabilitation became the leader in rating scale measurement and practice among health care publications. A key moment arrived when the Archives published an article that criticized the use of ordinal rating scale data as though they were interval measures (Merbitz, et al., 1989) and concluded that rating scale data were incapable of providing a basis for the scientific measurement of outcomes. Several letters to the editor pointed out the possibilities for an enhanced scientific basis for rating scales that exist in Rasch's models, and the editors invited Wright and Linacre to write a special article expanding on this theme (Wright & Linacre, 1989).

After the 1989 Wright and Linacre article, research employing Rasch models began appearing as articles in the Archives and other journals (a sampling of the articles at hand includes: Cella, Lloyd, & Wright, 1996; Chang & Chan, 1995; Daltroy, et al., 1992; Fisher, A., 1992, 1993; Fisher, W., 1993; Fisher & Fisher, 1993; Fisher, Harvey, & Kilgore, 1995; Fisher, Harvey, Taylor, et al., 1995; Granger & Wright, 1993; Grimby, et al., 1996; Haley & Ludlow, 1992a, 1992b; Haley, McHorney, & Ware, 1994; Heinemann, et al., 1994; Kilgore, Fisher, Silverstein, et al., 1993; Linacre, et al., 1994; Ludlow, Haley, & Gans, 1992; Lunz & Stahl, 1990, 1993; McArthur, Cohen, & Schandler, 1991; McHorney, Haley, & Ware, 1997; Pollack, Rheault, & Stoecker, 1996; Silverstein, Fisher, Kilgore, et al., 1992; Stucki, Daltroy, Katz, et al., 1996; Zhu & Cole, 1996), and not just as abstracts of annual meeting presentations. In 1991, a report on the Functional Independence Measure (FIM) employing Rasch models was made to the National Institute on Disability and Rehabilitation Research. The authors included Allen Heinemann, PhD, working at the Rehabilitation Institute of Chicago, and his colleagues Carl Granger, MD, and Byron Hamilton, PhD, of the Uniform Data System for Rehabilitation at the State University of New York in Buffalo, along with Wright and John Michael Linacre.

In 1993, the American Journal of Occupational Therapy published the proceedings of a 1991 conference sponsored by

the American Occupational Therapy Foundation and held at the University of Illinois-Chicago. Half of the papers elaborated on the scientific advantages of Rasch's models. Then in 1993, the journal Physical Medicine and Rehabilitation Clinics of North America published the proceedings of a 1992 conference hosted by Granger and Hamilton at SUNY-Buffalo; seven of the 13 articles were based on a Rasch analysis.

Since 1993, the research group at Marianjoy has moved to the Rehabilitation Foundation, Inc. (RFI), with Richard Smith in charge of the measurement and evaluation work. Also in the last five years, the number and type of journals in health care publishing Rasch analyses has grown considerably. The Journal of Clinical Epidemiology has published three articles in the last several years, and a research report (Campbell, Kolobe, Osten, et al., 1995) employing a Rasch analysis in Physical Therapy was nominated as "the article of the year."



Researchers at Wayne State University, American University, and Indiana University have developed significant work in outcome measurement for physical and health education, especially as these concern persons with disabilities (Spray, 1987, 1990; Safrit, Cohen, Costa, 1989; Safrit, Zhu, Costa, et al., 1992; Zhu & Safrit, 1993; Cole, Wood, & Dunn, 1991; Zhu, 1996; Zhu & Cole, 1996; Zhu & Kurz, 1994). Although this work situates itself within Item Response Theory, much of it, in fact, takes a strong measurement theory approach.

A MEDLINE search of the years 1993-1998 in the bibliographic database done in February, 1998, using the key word string, "Rasch analysis or Rasch measurement or Rasch model," produced 45 hits of articles appearing in 24 journals. Single articles have appeared in Stroke; Aging; Pain; Neurology; Arthritis Care and Research; Biometrics; and Nutrition & Health. Six articles appear in four Scandinavian journals, and one each in British, German, and French Canadian journals. The Archives of Physical Medicine & Rehabilitation has the most Rasch articles in the 1993–1998 period, with eight. The American Journal of Occupational Therapy and the American Journal of Physical Medicine & Rehabilitation both have five, with the Journal of Clinical Epidemiology and the Scandinavian Journal of Rehabilitation Medicine each with three.

The results of this search are limited to only what is included in the database. Not included was the 1997 special issue of Physical Medicine & Rehabilitation: State of the Art Reviews, edited by Richard Smith, which presents the proceedings of the First International Outcome Measurement Conference. Significant work in this area has also appeared in the Objective Measurement book series (Fisher, A., 1994; Ludlow & Haley, 1992; Ludlow & Haley, 1996; McArthur, Casey, Morrow, et al., 1992), as well as in non-medical journals, such as the International Journal of Educational Research (Fisher, A., et al., 1994).

To take advantage of Rasch's models for measurement we will need to establish the extent to which we can depend on these constructs as bases of comparison for the populations we serve. This calls for new ways of formulating research questions, reporting results, and collaborating, but most of all it requires a new awareness in the psychosocial sciences of the importance of metrology, the science of maintaining and improving the reference standard metrics through which we will most fully capitalize on scale-free measurement principles (Fisher, 1997a, 1997b, 1997c). For the latest on what's happening in the metrology movement among outcome measurement practitioners, be sure to attend the 2d International Outcome Measurement Conference at the University of Chicago, May 15-16.

Instantaneous Measurement and Diagnosis

John M. Linacre, Ph.D.
MESA Psychometric Laboratory
University of Chicago

The manufacture of measuring instruments is typically a large-scale, standards-based process. Their use is frequently an on-demand, local operation requiring immediate measures and measure interpretation. The FIM has been calibrated on large samples. These calibrations are used to construct the KeyFIM, a one-page data collection, measurement, and analysis device. This provides the physician the same measurement ease and immediacy as the yardstick does the carpenter. The KeyFIM incorporates the measurement replication and quality control diagnosis that the careful carpenter obtains by multiple measurements of the same unknown length.

Better Measurement

Better measuring instruments are not only more accurate and precise, they are also more immediate and intuitive. In industrial instrumentation, "better measurements, and more of them, have made it possible to interpret most data without recourse to statistical techniques" (Youden W.J., 1954).

Statistical techniques, particularly as implemented in computer programs, enable the calibration of observation instruments, such as the FIM, on large samples of patient records, representing many impairment groups and rehabilitation institutions. Collecting and analyzing large patient-record databases is an expensive and time-consuming process. Although this process yields useful information about the FIM and the patients to which it has been applied (Granger et al. 1993), it is far too slow and cumbersome to assist in the treatment of the patients whose records are in the database.

Effective use of the FIM requires that data collection, analysis, and interpretation occur almost instantaneously, preferably while the clinician is still with the patient (as with the clinical thermometer and stethoscope) or at least in a day or so (as with hospital-based laboratory tests). The increasing speed and ubiquity of computers will ultimately permit the development of artificially-intelligent systems to support the real-time analysis and interpretation of a patient's ratings on the 18 FIM items. Such interpretation will be based on the accumulated case histories of millions of patients to whom the FIM will have been administered. Nevertheless, the immediate local clinical experience of practitioners and their personal knowledge of the particular patient will always play a part in FIM interpretation.

Most of the benefits of a sophisticated computer-based system can be realized immediately with the KeyFIM, a simple, paper-and-pencil implementation of the FIM. This form combines into one graphical presentation the essential steps of data collection and measurement construction, along with a convenient layout for intuitive quality control and diagnostic interpretation.

Calibrating the Measurement System

The FIM consists of 18 items, each rated on a seven-category rating scale with each succeeding category carefully defined to represent an increasing degree of functional independence. It is designed to be administered to patients on admission to and discharge from a rehabilitation institution. Data collected from thousands of applications of the FIM have been subjected to extensive analysis. Linacre et al. (1994) report that analysis of FIM data from a measurement perspective by means of the Rasch model discloses that decomposing the 18-item FIM into 13 motor items and the 5 cognitive items produces two bases for measurement, clearly superior to the one composite original. For convenience, this paper focuses only on the FIM cognitive items, but the same considerations apply directly to the motor items.

Analysis of the FIM was conducted in the computationally convenient unit of measurement known as the Logit (log-odds unit, see Linacre, 1993, for other derivations). Though the Logit has a clear probabilistic interpretation (Wright & Stone, 1979 p. 36), its substantive interpretation depends on the use to which the measures are put. FIM measures are used in a rehabilitation setting in which clinicians expect patients to be functioning within a bounded

range of the conceptually infinitely wide variable (dimension, construct) of independence. The variable is infinitely wide, because it is always possible to imagine a patient even more dependent than any encountered to date (e.g., in a deeper coma), or even more independent than any encountered (e.g., with greater physical and mental prowess). The bounded range of independence is that for which the rehabilitation setting is designed. Accordingly, it is convenient to define a measurement scale with its "0" point corresponding conceptually to the lowest level of functioning at which a patient might be administered to rehabilitation. Similarly, the "100" point is defined to be the highest level of functioning which a patient might achieve and still remain in rehabilitation. In order to maintain the interval-scale measurement characteristics of the logit (Stevens, 1951), this "0" to "100" scale is a linear transformation of the logit scale. For clarity in substantive use, the new units of measurement are called FIMITs (Linacre, 1995).

FIM Cognitive Items		
Item Name		FIMIT calibration
N.	Auditory Comprehension	42
O.	Verbal Expression	40
P.	Social Interaction	46
Q.	Problem Solving	55
R.	Memory	52

Table 1. FIM Cognitive Items, condensed from FIM Guide (1993).

FIM Levels		
NO HELPER		FIMIT Step Calibration
7.	Complete Independence	24
6.	Modified Independence	8
HELPER		
5.	Supervision	1
4.	Minimal Assistance	-5
3.	Moderate Assistance	-11
2.	Maximal Assistance	-17
1.	Total Assistance	-

Table 2. FIM Rating Scale, condensed from FIM Guide (1993).

Expected Measures on FIM Cognitive Items								
Item Name	Level:	1	2	3	4	5	6	7
N.	Auditory Comprehension	8	24	34	41	49	61	82
O.	Verbal Expression	5	22	31	39	47	59	80
P.	Social Interaction	11	27	37	44	52	64	85
Q.	Problem Solving	20	37	46	53	61	73	94
R.	Memory	18	34	44	51	59	71	92

Table 3. Expected FIMIT measures for each Level on each FIM Cognitive Items.

Tables of corresponding values of FIM raw scores and FIM measures (in FIMITs) are given in Heinemann et al. (1994), as well as item calibrations in logits. For the purposes of constructing the KeyFIM, the Cognitive score-to-measure conversion table (op. cit., Table 4) was recomputed based on a random sample of 15,439 relevant patient records from the Uniform Data System (UDS) database using the BIGSTEPS computer program (Linacre & Wright, 1991). For the purposes of constructing the KeyFIM, a useful substantive range was obtained when the linear conversion is 12.5 FIMITS per logit. Table 1 contains FIMIT calibrations for the FIM item difficulties for this sample. Table 2 contains FIMIT calibrations for the adjacent category (step) calibrations. Table 3 contains the expected FIMIT measure corresponding to each possible rating on each FIM item. Since the expected measure for an extreme category is infinite, i.e., out of the operational range of the FIM, a Bayesian adjustment is made so that, for the extreme categories 1 and 7, the measures corresponding to expected FIM ratings of 1.25 and 6.75 are listed.

For most IGCs (except 1.1, 2, 12)		
FIM raw score on 5 cognitive items	FIMIT measure	FIMIT S.E.
5	0	17
6	8	12
7	17	9
8	22	7
9	25	6
10	28	6
11	30	5
12	32	5
13	34	5
14	36	5
15	38	5
16	40	4
17	41	4
18	43	4
19	44	4
20	46	4
21	47	4
22	49	4
23	51	5
24	52	5
25	54	5
26	56	5
27	58	5
28	61	6
29	63	6
30	67	6
31	70	7
32	75	8
33	81	10
34	91	13
35	100	18

Table 4. FIM raw scores to FIMIT measures conversion table.

Table 4 contains a FIM cognitive raw score to FIMIT measure conversion table. This covers most impairment group codes (IGCs), except groups 1.1 (left-hemisphere stroke), 2 (brain dysfunction), and 12 (congenital deformity).

Constructing the KeyFIM

The measures and calibrations presented in Tables 1-4 are sufficient to draw the KeyFIM shown in Figure 1. To explain its features and demonstrate its use, the analysis of two patient records is described here.

Figure 1. KeyFIM data collection and analysis sheet.

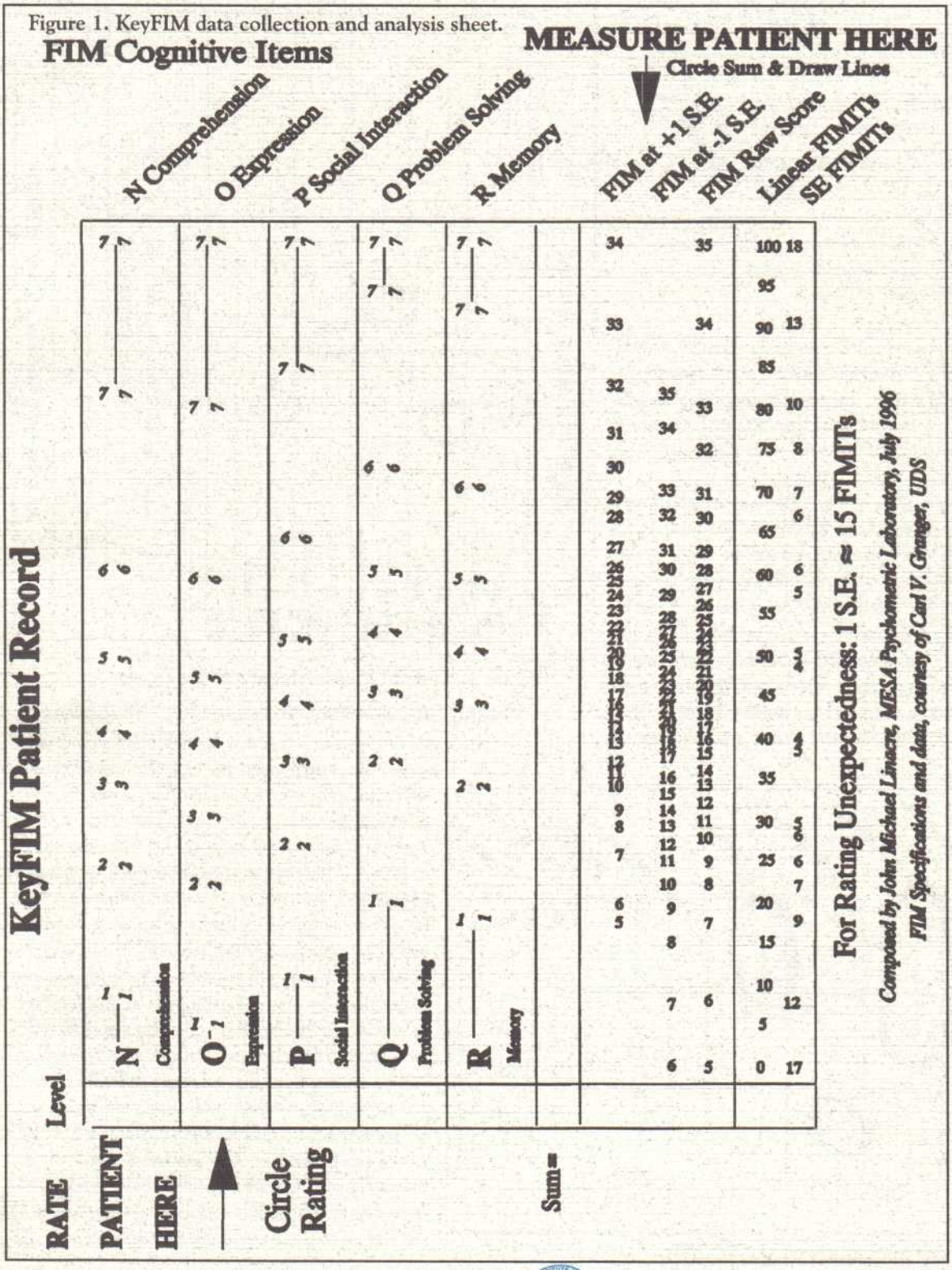


Figure 2 shows an actual patient record from IGC group 13, "Other Impairments." The KeyFIM has been turned on its side and the FIM levels recorded for each of the 5 cognitive items: 3 on item N. Comprehension, 3 on item O Expression, etc. The FIM ratings total 16. The corresponding levels are circled in the body of the KeyFIM. Data collection is now completed.

Figure 3 depicts the analysis stage. The KeyFIM is rotated, and a line drawn through the FIM raw score of 16 in each of three columns. The column "FIM at +1 S.E." indicates a high measure corresponding to one standard error of measurement above the expected measure. Continuing the line, by eye, to the "Linear FIMITs" column, indicates that a high measure corresponding to a raw score of 15 is about 45 FIMITs. The column, "FIM at -1S.E.," indicates a low measure one standard error below the expected measure. The "Linear FIMITs" column indicates that this is about 35 FIMITs. The third column, "FIM Raw Score," indicates that the expected measure for a score of 16 is about 40 FIMITs. The right-most column indicates that the standard error of this mea-

REHAB MEASUREMENT

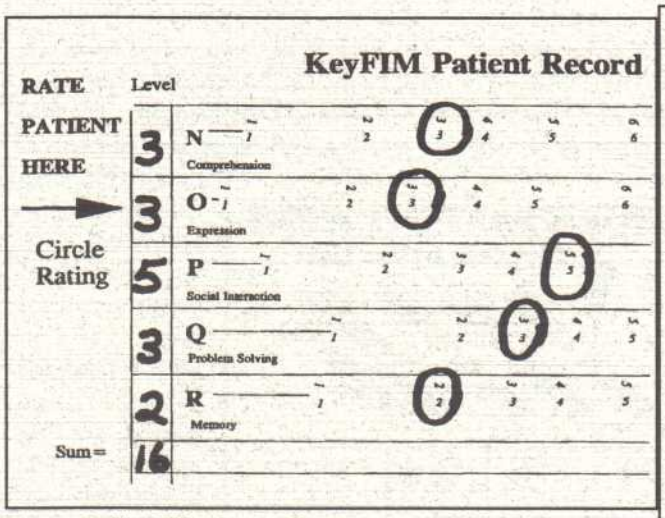


Figure 2. KeyFIM data collection.

sure is about 4 FIMITs, i.e., about the range 35-45 as illustrated. The legend on the right of the Figure states "For Rating Unexpectedness: 1 S.E. = 15 FIMITs." Based on conventional statistical testing, observations located further than 30 FIMITs from the mean line would be suspect, but here the most outlying, "5" on Social Interaction, is only 15 FIMITs away.

In this example there are no observations in extreme categories, but these require special treatment. A rating at an extreme level "1" or "7" corresponds to an infinite range of performance away from the next most extreme category. Accordingly, this is shown by a "—" on the KeyFIM. Thus for "7" on N. Comprehension, the KeyFIM has "7—7." This means that any location along the "—" is a reasonable location for the rating to be marked on the form. In practice, ring around the entire region, as in Figure 4, and choose the point on the line most consistent with the other ratings for measurement and fit analysis purposes.

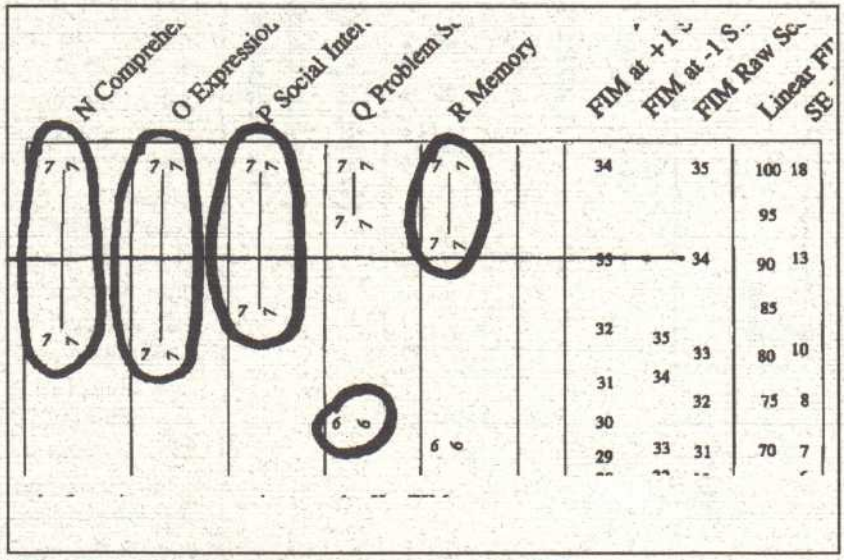


Figure 4. Locating extreme ratings on the KeyFIM.

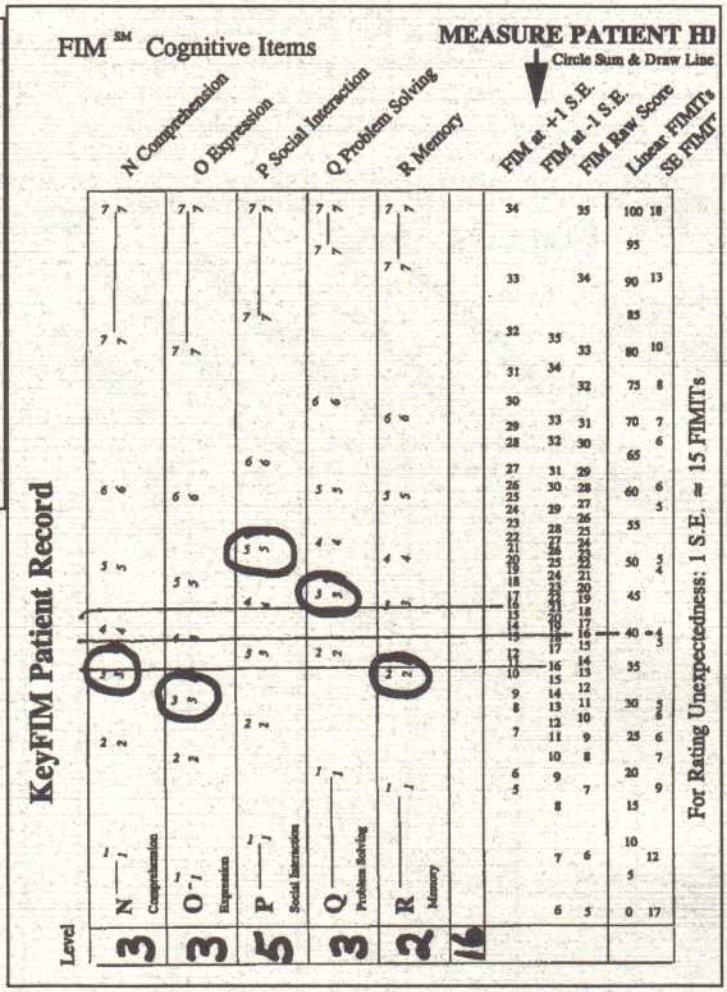


Figure 3. KeyFIM Measurement and Fit Diagnosis.

Instantaneous Measurement and Diagnosis

Since each FIM item provides a locally independent measure of functional independence, they can be used as the basis for an intuitive, rather than statistical, measurement process. Figure 5 provides an example of another actual patient record. Here the observation to item R. Memory has been deliberately omitted — as though it were not yet recorded, perhaps never to be. There is no "complete" raw score, so the horizontal lines cannot be drawn directly. Intuitively, it is clear that the patient's typical level of independence is described by the higher ratings. A line has been drawn by eye through these, yielding a general independence of 58 FIMITs. The S.E. of this measure will be greater than the indicated 5 FIMITs due to the missing observation and discrepant rating pattern, treating the precision of this measure as 8 FIMITs would be reasonable. The low rating of "2" on Expression is at 20 FIMITs, about 38 FIMITs below the typical level. 38 FIMITs is twice the rating S.E. of 15 FIMITs, so that this rating is statistically unexpected.

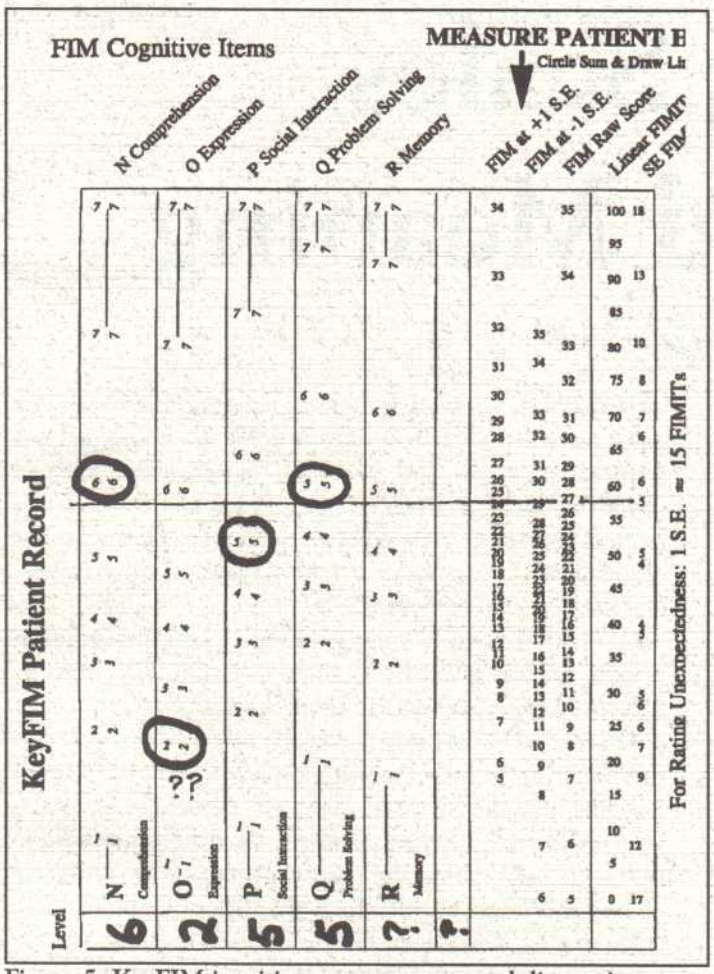


Figure 5. KeyFIM intuitive measurement and diagnosis.

More important for practice, however, is that it is obviously an outlier according to Leonard "Jimmy" Savage's "intra-ocular traumatic test." For clinical practice, it is this rating that will motivate the patient's immediate therapy.

In this example, measurement and fit diagnosis proceeded successfully and immediately despite incomplete data and the inability to use a "complete" raw score as the basis of analysis. Further, fit analysis and diagnosis could have proceeded successfully even without any formal statistical tests.

Conclusion

The KeyFIM is an example of how any rating instrument can be presented as a self-measuring form, supporting intuitive measurement and fit diagnosis. Its format encourages the practitioner to evaluate the ratings as they are being collected, so avoiding obvious data entry errors and misunderstandings. With a little experience, the practitioner can perform measurement and fit analysis in the same immediate, effortless and routine way that useful measurements are obtained from bathroom scales and clinical thermometers. The KeyFIM and instruments like it further blur the artificial distinction between physical and psychological measurement.

Notes:

1. "The term accuracy usually denotes in some sense the closeness of the measured values to the true value, taking into consideration both precision and bias. Bias [is] defined as the difference between the limiting mean [of observations] and the true value" (Ku H.H. 1967). See also "Use of the Terms Precision and Accuracy as Applied to the Measurement of a Property of Material" (ASTM Designation, E177-61T, 1961).

2. The Bayesian adjustment for extreme scores and ratings employs this line of reasoning: the KeyFIM would not have been administered to the patient if there were no chance that the patient might have been observed in a non-extreme category. Accordingly, the observation in the extreme category was barely enough to qualify as extreme. For extreme scores, this corresponds to an unobservable raw score that is 0.5 raw score points away from the extreme, i.e., a raw score of 35 out of 35 is treated as a score of 34.5, and a raw score of 5 out of 35 is treated as a score of 5.5. For individual ratings, performances in the range 1.5 to 2.5 would be observed as ratings of level "2." Ratings less than 1.5 would be observed in the extreme level of "1." Consequently any performance from 1 to 1.5 is observed as "1," and a "1" is treated as an "average" rating of 1.25 for the purposes of locating the category on the KeyFIM. Similarly, a "7" is treated as a 6.75. References:

Granger C.V., Hamilton B.B., Linacre J.M. Heinemann A.W., Wright B.D. (1993) Performance profiles of the Functional Independence Measure. American Journal of Physical Medicine and Rehabilitation 72:2 April 84-89.

FIM Guide (1993) Guide for the Uniform Data Set for Medical Rehabilitation (Adult FIM). Version 4.0. Buffalo, New York: State University of New York at Buffalo.

Heinemann A.W., Linacre J.M., Wright B.D., Hamilton B., Granger C.V. (1994) Measurement characteristics of the Functional Independence Measure. Topics in Stroke Rehabilitation 1(3) p.1-15. Fall.

Ku H.H. (1967) Statistical Concepts in Metrology. Chapter 2 in Handbook of Industrial Metrology. American Society of Tool and Manufacturing Engineers. p. 20-50. New York: Prentice-Hall.

Linacre J.M. (1993) Why logistic ogive and not autocatalytic curve? Rasch Measurement Transactions 6:4 p. 260-261.

Linacre J.M. (1995) KeyFIM -Self-Measuring Score Form. Rasch Measurement Transactions 9:3 p. 453-4.

Linacre J.M., Heinemann A.W., Wright B.D., Granger C.V., Hamilton B.D. (1994) The structure and stability of the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation 75(2), 127-132, Feb.

Linacre J.M., Wright B.D. (1991) BIGSTEPS Rasch measurement computer program. Chicago: MESA Press.

Stevens S.S. (Ed). (1951) Handbook of experimental psychology. New York: Wiley

Wright B.D., and Stone M.H. (1979) Best Test Design. Chicago IL.: MESA Press.

Youden W.H. (1954) Instrumental Drift. Science 120:3121 p. 627-631. October 22, 1954.

John Michael (Mike) Linacre, Ph.D., M.A., C.D.P., C.C.P.

Dr. Linacre is Associate Director of the Measurement, Evaluation and Statistical Analysis (MESA) Psychometric Laboratory at the University of Chicago. After obtaining a degree in Mathematics from Cambridge University in 1967, he engaged in computer-related activities in England, Japan, Australia, and the USA. In 1981, he worked with Prof. Benjamin Wright to develop the Rasch analysis computer program, Microscale. In 1986, Mike moved to the University of Chicago and obtained a Ph.D. in psychometrics. Since then he has conducted research, taught classes and continued the development of Rasch computer programs, most recently Facets and WINSTEPS.

Rating Scales and Shared Meaning

Winifred A. Lopez, Ph.D.

A rating scale is an aid to disciplined dialogue. Its precisely defined format focuses the conversation between the respondent and the questionnaire on the relevant areas. All respondents are invited to communicate in the shared language of the specified option choices (Low 1988).

Ambiguity and uncertainty, however, remain. First, some respondents may not use the rating scale as it was intended to be used. Choosing socially acceptable responses or falling into a response set defeats the purpose of the questionnaire. Second, respondents can only interpret a rating scale in terms of their own understandings of category labels. Lack of clear, shared category definitions invites ambiguity and idiosyncratic category use. Different interpretations lead to inconsistent use patterns.

Traditional statistical analysis, however, mistreats all rating scale observations as precise and accurate communications. Researchers seldom provide for differences in perspectives among respondents. These differences cannot be overlooked if our objective is the pursuit of useful knowledge and sound decision-making. We must recognize the various ways in which rating scale categories might be used and identify those which enable the maximum extraction of meaning. While this involves choice on the part of the analyst, "selective emphasis, choice, is inevitable whenever reflection occurs" (Dewey 1925). Because there can be no knowledge without choice, it becomes the responsibility of the analyst to develop criteria by which those choices can be made.

"Meanings do not come into being without language and language implies two selves in a conjoint or shared understanding" (Dewey 1925). Some level of ambiguity is unavoidable because language can never be exact. Nevertheless, shared meaning cannot be extracted from individual responses unless analysis can identify a common, cooperative mode of communication among all parties concerned.

Rating scale analysis must take the perspective that while a rating scale offers respondents a common language, a tool for "categorizing, ordering and representing the world" (Halliday 1969), it does not by itself make for meaningful communication. Since "meaning is located neither in the text nor in the reader but in their interaction" (Bloome & Green, 1984), we must include a step concerned with discovering, rather than asserting, meaning as we conduct our statistical analyses. Just as readers "must choose between competing interpretations of text" (Bloome & Green 1984) so must the analyst choose between different interpretations of the rating scale in order to find a coherent, shared representation of what is investigated.

A rating scale, like any other tool, "is defined by how it is used" (Halliday 1969). A focus of our analysis must be how the rating scale is actually used by respondents. We must discover which transformation of the initial rating scale categorization extracts the "maximum amount of useful [shared] meaning from the responses observed" (Wright et al. 1992).

As shared meaning develops, we establish criteria so that we do not ignore the individual, but rather provide a scoring medium through which the dissenting individual's voice may be heard more clearly. We set the stage so that individuals who do not subscribe to our construction of shared meaning can stand out and be noticed. By establishing an explicit commonality among most respondents, we enable the meaning which stems from an individual's unique interaction with an item or a group of items to emerge.

The constructive analysis of rating scale data can promote both general dialogue with the group and specific dialogue with the individual.

Bloome D, Green G (1984) Directions in the sociolinguistic theory of reading. In PD Pearson (Ed.), Handbook of Reading Research (pp 395-421). White Plains NY: Longman.

Dewey J (1925). Experience and nature. Republished in J.A. Boydston (Ed.) John Dewey: The Later Works, 19925-1953, Vol. 1. 1981. Carbondale IL: Southern Illinois University Press.

Halliday M (1969) Relevant models of language. Educational Review, 22, 1-128.

Low GD (1988) The semantics of questionnaire rating scales. Evaluation and Research in Education 2(2), 69-70.

Wright BD, Linacre JM (1992) Combining and splitting categories. RMT 6:3, 233.

Rating Scale Categories:
Dichotomy, Double Dichotomy, and the Number Two

Mark H. Stone, Ph.D.
Adler School of Professional Psychology
Chicago, IL

My conjecture is that dichotomies in rating scales are more useful than multiple ratings. This conjecture implies that most multiple ratings can be reduced to a useful natural dichotomy making construction of multiple ratings futile. Why do I maintain such a conjecture when most rating scale practice uses multiple categories?

Personality Inventories

First, I illustrate my point by reminding the reader that the most utilized of all standard personality inventories is the Minnesota Multiphasic Personality Inventory, the famous MMPI. It uses a dichotomy, true/false, for response alternatives. MMPI protocol allows a "?" or "cannot say" response as an alternative. But the directions ask the test administrator to encourage the respondent to return again to such responses and to decide in which direction to mark the answer. The goal is to eliminate middle responses.

I don't argue that just because the test authors recommend this, it is correct. I only remind the reader that if multiple ratings had been found to be more advantageous, you can bet they would appear on the test protocol. I suggest this has not occurred because after decades of use the dichotomy still works.

Indeed, multiple ratings have not been found to add information, but rather provoke noise. When the number of "?" responses is high it is a sign that the validity of the entire test is in question. Graham (1987, p. 19) says, "... the validity of a resulting protocol with many omitted items should be questioned..." and "... encourage individuals to try to answer previously omitted items, most people will complete all or most of the items." Graham says the same in his text on the revised edition MMPI2 (Graham, 1993). The MMPI Manuals for both editions recommend the same procedures.

We see that "forcing" a dichotomy is standard administrative practice for the two editions of the MMPI and the same can be said for the competing personality inventories, the Millon Clinical Multiphasic Inventory, California Personality Inventory and 16PF.

The earliest edition of the MMPI produced each item separately, printed on a card, and the patient placed cards sorted "true" in one box and those sorted "false" in the other. I have always considered this process an intelligent procedure for patients inasmuch as most of the people taking the MMPI are less then optimally functional. Any strategy that assists them ought to be promoted. Sorting is a tactile activity as well as a cognitive one that is advantageous to the subject. It gives the respondent the opportunity to "handle" the question and physically sort as opposed to marking responses with a pencil on an answer sheet. The size of the response window has progressively decreased over the years. I doubt this has brought much advantage to respondents. The main impetus for answer sheets is that the original sorting routine was troublesome to score for psychologists. Today's streamlined answer sheet can be quickly scanned. Good for psychologists. Bad for subjects.

From appraising well-known personality inventories, we observe that patients are asked to make dichotomous decisions to each item. To inquire into motivation and other confounding variables behind their responses takes us away from the problem at hand, but requiring a true/false or yes/no response clearly seems the most useful way to collect responses from patients. For people under stress, this is the most reasonable expectation and solution. Of course, personality inventories are not rating scales, but the problem of determining a valid response alternative is common to Likert scales and personality inventories, and the latter have promoted the dichotomy for more than 50 years with little motivation to change.

I think this example adds support to my conjecture, but taken alone it is not an overwhelming argument for advocating a dichotomy. What adds more evidence to my conjecture comes from the reasoning of individuals about the status of a dichotomy in general. There are several quotes worth thinking about.

Karl Menninger in his book on Number Words and Number Symbols says

> *"Two has a special status and is not just a number like any other in the number sequence, but instead is that extra ordinary number ... ."*

He then goes on to say that the number two has more significance then we might assume today in the era of big numbers. It occupies a unique place after "one." But it is not only the second numeral in our counting system. Two suggests something beyond "one more" because at this juncture we enter upon the idea of contrasts, comparisons, and opposites.

The proverbial essay question that teachers frequently give to students often requires "contrast and compare" in some form or another. We pursue many tasks efficiently and effectively by dichotomous grouping, particularly when they are vo-

luminous and tend to overwhelm us. Consider the following categories and just one of the dichotomous groupings that can result.

Spelling: words spelled phonetically vs. words that are not.
Grammar: regular verbs vs. irregular ones.
Math: plane geometry vs. spherical.

Alfred Adler and other psychologists have suggested that a dichotomy is generally the haven of the perplexed, the neurotic, and the primitive mind. The dichotomy comes forth whenever we feel pressured or at risk. At such times we formulate response alternatives by a dichotomy, not by imagining an array of alternatives. So whenever respondents do not know how to answer an item they respond by falling back on a dichotomy.

Jung also thought two had a special value.

"Two is the first number because, with it, separation and multiplication begin, which alone make counting possible."

What the number two brings us is a phenomenon that is omnipresent:

- from the body: two eyes, two ears, two hands, two feet, two kidneys, two lungs.
- from nature: male/female, night/day, sun/moon.
- from contrasts: old/young, right/left, up/down, plus/minus, hit/miss.
- from mythology: god/goddess, two in one — twins, the Egyptian double lion, named Routi.

Given this ubiquity for opposites, are we not more attuned to a dichotomy than to any other system?

Edward Edinger (1995) expands Jung's point in discussing Moby Dick.

A major theme of Moby-Dick is the problem of opposites. As we proceed we shall encounter numerous antitheses: alienation and inflation, courage and cowardice, strength and weakness, black and white, good and evil, the bounded land and the boundless sea, height and depth, the universal and the particular, Christian and pagan, primitive and civilized, the outer word and the inner soul, spirit and matter, destiny and free will, love and hate, calm and turbulence, delight and woe, orthodox and heretic, reason and madness, God and man. (p. 30)

Paul Tillich, the philosopher/theologian adds this point,

"Philosophical ideas necessarily appear in pairs of contrasting concepts, like subject and object, ideal and real, rational and irrational."

Tillich reminds us that ideas are "paired," that for every point we conjecture an opposite.

Lastly, C.S. Peirce, the American philosopher/logician expresses in a more comprehensive view the totality of what is found in the first three numbers.

"First is the conception of being or existing independent of anything else. Second is the conception of being relative to, the conception of reaction with something else. Third is the conception of mediation. ... The origin of things, considered not as leading to anything, but in itself, contains the idea of First, the end of things that of Second, the process mediating between them that of Third."

What these thinkers have to say about "two-ness" and the dichotomy is more than idle speculation. They are speaking about a phenomenon that permeates our thinking about the number two and a dichotomy. We see most concepts in terms of dichotomies — pairs, opposites, and contrasts.

George Miller (1956, p. 82) offers commentary that is relevant in his paper entitled, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Miller defines "amount of information" as variance which is a dimensionless quantity. He goes on to say,

"When we have a large variance, we are very ignorant about what is going to happen. If we are very ignorant, then when we make the observation we get a lot of information. On the other hand, if the variance is very small, we know in advance how our observation must come out, so we get little information from making the observation." (p. 82)

The key point from Miller which applies to rating scales is whether or not we "get a lot of information." This can only occur with multiple ratings when a two-step model is shown empirically to be more informative than a one-step model, and threesteps is shown to be more informative than two steps. Instead, the construction style for most Likert scales seems to be slapping as wide a range of response alternatives as possible to a varied collection of poorly worded items. Such a process cannot produce information.

From this state of ignorance it is possible to "collect data," but the quality of such responses is unknown and suspect. Not knowing how a person will answer an item is an entirely different problem from not knowing what the possible response alternatives might mean to a range of respondents. In the former situation we have the state of ignorance prior to knowing the outcome. In the second situation we are simply ignorant of how to build a response alternative that is meaningful. We might want to read the thermometer with scientific dispassion, but we do not construct a thermometer dispassionately! We give its construction our best attention. There is a big difference between these two states of ignorance, and there appears to be misplaced credence in believing that "ignorance" expresses the desired state of neutrality in scientific work. If we propound ignorance do we produce knowledge or only become more confused?

There is one response scheme that is popular on rating scales. It builds on a double dichotomy of four alternatives. A common example is "Strongly Agree, Agree, Disagree or Strongly Disagree." Miller (1956) informs us that "Two bits enables us to decide among four equally likely alternatives" (p. 83). As the number of alternatives increases by a factor of

two, one more bit of information is added. Consequently, eight alternatives equals three bits, which is about as many response alternatives as are ever found on a rating scale.

Miller says,

> *"It is interesting to consider that psychologists have been using seven-point rating scales for a long time, on the intuitive basis that trying to rate into finer categories does not really add much to the usefulness of the ratings" p. 84.*

He goes on to cite four experiments in which a good observer can identify about four intensities, about five durations, and about seven locations. Miller argues that our nervous system gives us a finite limit to our capacity for making judgments. This limitation does not vary much from one sensory attribute to another.

His article concludes by saying (1) we have definite limitations of absolute judgment (2) chunking helps and is the only way we can address this limitation. In his summary, Miller suggests,

> *"the recoding that people do seems to me to be the very lifeblood of the thought processes" p. 95.*

With four alternatives we must solve two dichotomies. The first one is 1-2 vs. 3-4 followed by deciding between 1-2 or 3-4, or else the item is resolved as a single dichotomy 1 vs. 2-4 or 1-3 vs. 4. We solve a double dichotomy of four responses by chunking the problem into two groupings of two each — two successive dichotomies — or else form it into a single dichotomy.

Lastly, Miller proffers his theory as

> *"a yardstick for calibrating our stimulus material and for measuring the performance of our subjects" p. 96.*

His conclusion of a natural limit of three bits makes eight alternatives the maximum according to his evaluation of four physiological and memory studies. He concludes that the practical span of alternatives is, in fact, much smaller than eight. On the basis of his studies we are advised to reduce rather than expand the number of ratings. Miller infers that through chunking and recoding we resolve a large number of alternatives into a smaller number. The process may occur so quickly with some items as to make us think it is a single solution, but whenever we have to pause and deliberate over multiple ratings, it is clear that chunking and regrouping are operating.

We need to be aware of the limitations of our nervous system and not offer the possibility of multiple ratings when, in fact, they are not easy to resolve. Multiple ratings have to be demonstrated as empirically operating, not imagined to do so. It is doubtful that we can actually cope systematically with many alternatives. What we learn from Miller's investigations is that the dichotomy is not easily transcended.

Support for my conjecture of the dichotomy also comes from considering the practice of rescoring response alternatives. I present two examples, showing in both of them that the rescoring of four alternatives is efficiently reduced to two.

The first example concerns the Beck Depression Inventory. It was administered in the Adler clinic to 266 non-clinical subjects and 153 clinically depressed persons. This scale of 21 items has four responses to each item indicated by 0, 1, 2, or 3. James Natter and I recoded these responses to a dichotomy of 0 = 0 and 1 = 1, 2, 3 which produced a dichotomous scoring model that differentiated between clinical depressed and non-clinical subjects better than the original category scale. A second rescoring dichotomy 0 = 0, 1 and 1 = 2, 3 was not as discriminating as the first, but still better than the original scale. The first dichotomy also produced better differentiation between persons attempting suicide or not in the depressed sample than did the original scale. Natter (1994) concluded that the original BDI scale is less effective than the dichotomy for differentiating pathology.

The second sample includes responses of 233 outpatient subjects in the Adler clinic taking the Wolpe-Lange Fear Survey Schedule II (1969). This scale is a self-report list of 108 items to which respondents endorse the amount of unpleasant feelings associated with each. **Table 1** gives the complete rescoring analysis for each of the 15 models.

Column 1 gives the scoring code.
Column 2 gives the steps in the model.
Column 3 PSEPR is the person separation reliability.
Column 4 PSEP is the person separation index.
Column 5 ISEP is the item separation index.
Column 6 UCON is the number of iterations for convergence.
Column 7 PINSD is the person infit standard deviation.
Column 8 IINSD is the item infit standard deviation.
Column 9 is the number of items identified beyond a standardized misfit of 2.0.
Column 10 ISEPR is the item separation reliability.
Column 11 PSEP/PINSD is the ratio of person separation to the person infit standard deviation.
Column 12 ISEP/IINSD is the ratio of item separation to the item infit standard deviation.

Examination of the results shows that model 01111, a one-step model, and model 01122, a two-step model, were better than the original model 01234, a four-step model. Model 01222 does better than any other two, three, or four-step models in ISEP and PSEP, but does produce misfit in 21 of the 108 items. Model 01111, however, while losing some ISEP and PSEP saves 12 of these items. This model is efficient. The ISEP and PSEP indices are among the highest values for several models. The number of fit items, although not the lowest, is less than eleven other models. Model 01111 contains only one step and indicates that the FSS can be efficiently scored as dichotomous. Columns 11 and 12 produce their highest values for the dichotomous model.

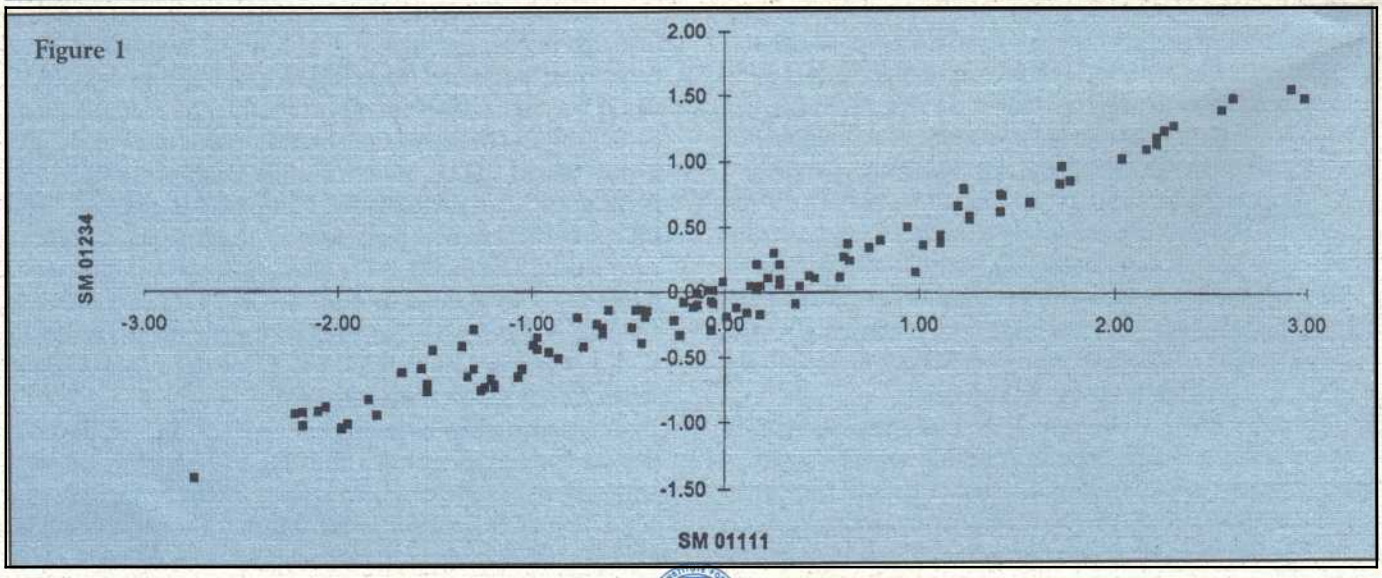
Comparing the dichotomous model of 01111 to the two-step model 01222 produces a PSEP ratio of 5.3/6.0 = .88 indicating the dichotomous model is 100(5.3/6.0) = 88% efficient of the best scoring model of the fifteen. The original model 01234 is 100(5.5/6.0) = 91% of model 01111, but at the cost

Table 1

Wolpe Fear Scale
Scoring Models Analysis
Mark Stone

	1	2	3	4	5	6	7	8	9	10	11	12
	Scoring Model	Steps in Model	PSEPR	PSEP	ISEP	UCON # Its	PINSD	IINSD	# Items Out	ISEPR	PSEP/ PINSD	ISEP/ IINSD
			[T. 3.1]	[T. 3.1]	[T. 3.1]	[T. 0.2]	[T. 3.1]	[T. 3.1]	[T. 3.1]	[T. 3.1]	C4/C7	C5/C8
1	SM 00001	1	0.71	1.6	1.9	4	0.07	0.13	4	0.78	22.86	14.62
2	SM 00011	1	0.87	2.6	3.3	4	0.12	0.12	8	0.92	21.67	27.50
3	SM 00012	2	0.84	2.3	3.2	8	0.27	0.17	8	0.91	8.52	18.82
4	SM 00111	1	0.94	4.0	5.4	3	0.16	0.11	11	0.97	25.00	49.09
5	SM 00112	2	0.94	3.8	5.2	3	0.30	0.16	11	0.96	12.67	32.50
6	SM 00122	2	0.93	3.8	5.2	8	0.25	0.13	10	0.96	15.20	40.00
7	SM 00123	3	0.93	3.6	5.0	14	0.36	0.19	14	0.96	10.00	26.32
8	SM 01111	1	0.97	5.3	7.4	4	0.17	0.09	9	0.98	31.18	82.22
9	SM 01112	2	0.97	5.5	7.4	20	0.40	0.18	15	0.98	13.75	41.11
10	SM 01122	2	0.97	5.8	7.5	9	0.35	0.18	20	0.98	16.57	41.67
11	SM 01123	3	0.97	5.4	7.2	10	0.51	0.25	26	0.98	10.59	28.80
12	SM 01222	2	0.97	6.0	7.9	7	0.30	0.14	21	0.98	20.00	56.43
13	SM 01223	3	0.97	5.7	7.7	4	0.44	0.19	21	0.98	12.95	40.53
14	SM 01233	3	0.97	5.7	7.6	10	0.40	0.19	22	0.98	14.25	40.00
15	SM 01234	4	0.97	5.5	7.4	15	0.51	0.25	22	0.98	10.78	29.60

Figure 1



of 15 items! A plot of the dichotomy vs. the original model in **Figure 1** shows that the person measures for the two models are consistent (r = .98). Simple identification of fear is sufficient. Attempts to discriminate further are not useful. The FSS functions very well when scored as a dichotomy.

In both examples, reduction to a dichotomy was a reasonable alternative. This is useful to know before beginning further study of the data. It might prove useful to retain the original format for administration, but it is clear that the original model is only a conjecture of what the authors imagined, not what occurred.

Conclusions

1. Personality measurement has employed the dichotomy for more than 50 years as a response alternative. The dichotomy has worked in this field.

2. The dichotomy is a fundamental phenomenon of mind according to those who have given it thought — Menninger, Adler, Jung, Edinger, Tillich, and Peirce. It operates most noticeably when we are overwhelmed by experience that needs reduction.

3. Miller explains that multiple ratings are limited by our span of comprehension and that we reduce multiple indices by chunking and regrouping, especially when overwhelmed.

4. Scoring model analysis indicates that dichotomous models are as good or better than some original scales. These examples show that devising multiple ratings requires more than attaching a rating scheme to an item.

Based upon my conjecture, here are some suggestions for scale construction.

1. Put yourself in the respondents' role and carefully determine what their responses might be. Utilize the psychology of human behavior to determine how respondents might behave. Don't just slap a rating scale to an item.

2. Write a strategic number of carefully crafted items that contribute to the construction of a unidimensional variable. Don't ask every question you can imagine.

3. Begin with a dichotomy and forget about having multiple ratings until a well-defined variable has been constructed. If you think it might be useful, expand the rating alternatives and evaluate the results.

4. Analyze all the scoring models to see how each is working. Don't begin with a rating scale and pretend it works.

To construct a good scale we first need to address the intent of the scale regarding person response behaviors, not write items. We need to identify the characteristics of the intended respondents. This will guide how items should be written and response model alternatives.

The major question is, "Do author and respondent models coincide?" If we do not make a careful analysis of responses we will never know the answer to this question. Many researchers accept the responses according to their intention without bothering to make an analysis of respondent behaviors. Scoring models need to be evaluated to determine how respondents view the scale. My conjecture suggests we should accept the preeminence of the dichotomy as the operating model until other alternatives can be demonstrated.

Ben Wright has suggested that the scale is a "conversation" between the author(s) of the scale and the respondent(s). This is a useful model for scale construction and it reiterates the idea that the first task in scale construction is not to write items, but to address the possible range of relevant person behaviors that could occur. I have suggested a number of steps to follow in item construction, but want to emphasize that planning for respondent behavior should always precede item writing.

The next step is creating a response format. I argue that rather than create the typical Likert response format, use a dichotomy to investigate whether a variable has been achieved. When a variable has been successfully constructed, investigate whether or not the measures are enhanced by a more complex scoring format. Proceeding in a step-by-step approach is more sensible than beginning with a more complex response scheme that may not work. When in doubt, keep it simple. Use a dichotomy.

References

Edinger, E. (1995). Melville's Moby-Dick: A Jungian commentary. Toronto: Inner City Books.

Graham, J. (1987). The MMPI: A practical guide. (2nd Ed.), New York: Oxford.

Jung, C. (1938). Psychology and religion. Volume 11. Collected Works of C.G. Jung. Princeton, NJ: Princeton University Press.

Linacre, J. & Wright, B. (1997). BIGSTEPS. Chicago: MESA Press.

Menninger, K. (1969). Number words and number symbols. New York: Dover.

Natter, J. (1996). Psychometric properties of the Beck Depression Inventory. Unpublished dissertation.

Peirce, C. (1940). in Philosophical writings of Peirce. J. Buchler (Ed.), New York: Dover.

Tillich, P. (1962). Existentialism and psychotherapy in psychoanalysis. in Existential Philosophy H. M. Ruitenbeak (Ed.), New York: Dutton.

There are three stages to the life of revolutionary scientific ideas. They are initially rejected as outrageous heresies, then they are recognized as brilliant discoveries, and finally they are assumed to be the way things have always been.

William James (paraphrased)

Measure Accuracy:
Functioning-Level vs. Grade-Level Testing

George S. Ingebo, Ph.D.

In grade-level testing, all grade-three students take the same grade-three test; grade-four students the grade-four test.

In functioning-level testing, students take tests designed for their attainment level, whether they are low-middle or high-achieving students. There may be five to six achievement levels at a grade-level level.

Functioning-level testing is not equivalent to out-of-level testing, in which low- or high-achieving students are tested at a lower or higher grade-level.

Data from test publishers' grade-level tests indicate that few grade-level test items accurately measure low- or high-achieving students' ability. This study shows that, when compared to testing students at grade-level, testing students at their functioning-level substantially reduces measurement error.

Since 1979 to the present, the Portland Public Schools of Portland, Oregon administers a basic skills testing system using functioning-level tests. Portland calibrates this testing system with a Rasch measurement model and maintains records on the performance of the students taking these tests at every grade during that entire time. Students take one achievement test out of a series of tests in the fall, and another in the spring.

Portland Public Schools selects a test level for a student by finding each student's score on their last district test. Portland places a student at an ability level based on their expected growth. By fitting a test to a student's established ability, whether it is a high- or low-achieving student, sufficient items in each functioning level test measure the performance of that student.

The State of Oregon State Assessment Program tests at grade-level rather than functioning-level. This study compares the Portland Public Schools functional-level testing with the State of Oregon grade-level testing results.

The State of Oregon employs the same testing procedures and the same Rasch scale used in the Portland Public Schools districts. The Oregon State and the Portland Public Schools have the same curricular goals in reading and mathematics. Both Portland Public Schools and Oregon State testing systems calibrate their tests for content difficulty with the same Rasch scaling model. These factors facilitate direct comparisons.

The State of Oregon administers grade-level tests to students once a year, in the spring. There are two state tests for mathematics and two for reading. Depending on where they live, students take one each of these state tests. The Portland Public Schools administers level-tests in fall and spring.

Procedure

We generate a quantitative probability that, given a test scaled in calibrated measures, each item in that test has a level of difficulty based on this scale. We predict the percentage of students who we expect to answer each item correctly from the calibrated item measures.

> *We expect that 50 percent of students who have attained an ability measure of 200 on previous tests to correctly answer an item that has a difficulty level calibrated at 200. We expect 25 percent of students with an ability level of 190 to correctly answer an item with a difficulty level calibrated at 200. We expect 75 percent of students with an ability level of 210 to correctly answer an item with a difficulty level calibrated at 200.*

Then, we estimate differences between expected and actual performance of every item of each student group achieving the same Rasch scale total score. We group students from low-, middle-, or high-achieving, based on their past ability measures.

We define test accuracy based on the amount, not the number, of deviations from the expected score. Tests with greater deviation amounts are less accurate. Tests with the deviation closer to expectation are more accurate.

Grade-level tests data were from 1993-94 state grade-five mathematics tests administered to two groups of students and another two groups of students taking grade-five reading tests.

Functioning-level tests data were from only one of five 1993 Portland Schools levels-tests administered to grade-five students.

We compute Rasch scale measures for all Oregon State

TEACHING

Assessment grade-five 1993-94 tests and the Portland spring 1994 functioning-level test scores. We exclude records of students getting less than 30 percent items correct from the analysis.

For each student scale score, we recover the probability of success for that student on each item in the test taken.

For all students getting the same Rasch scale measure in each compared group, we compute the differences between expected and actual performance.

We examine the differences between mean expected scores and the mean actual scores on all test items completed by each total-score group.

We examine the differences between the expected standard deviation and the actual standard deviation on all items attempted by students at each raw score level.

We aggregate the differences between expected and actual performance all items for all tests (reading and math) per increasing student measures.

Findings

1. For Oregon State grade-level tests and the Portland functioning-level tests, students had the same rate of number of differences. In both student groups, a similar ratio did better or worse than expected, Table 1.

2. The amount of difference between expected and actual performance is twice to three times as great for low-achieving reading students taking State of Oregon grade-level tests, Figures 1b and c, as for those taking the Portland schools functioning-level tests, Figure 1a.

3. The difference between expected and actual performance for low-achieving mathematics students is twice to three times more on the State of Oregon grade-level tests, Figures 1e and f, than the Portland functioning-level tests, Figure 1d.

4. Reading and math standard deviations aggregates show functioning-level tests are two times more accurate than grade-level tests, Figure 2.

5. Basic skills measures for all students are best for students with mid-range scores for grade-level and functioning-level tests. Differences in measurement accuracy between the grade-level and the functioning-level groups are less pronounced at the upper score levels than at the low score levels.

a) Oregon State Grade-level Tests

Grade-level Tests		Number of Students	Number of Differences	Items
Reading	351	7,545	1,064	38
	352	7,643	1,044	36
Math	451	7,443	1,380	46
	452	7,347	1,334	46
total		29,978	4,822	166

vs. b) Portland Area Functioning-Level Tests

Functioning-level Tests		Number of Students	Number of Differences	Items
Reading	504	7,512	1,000	40
Math	516	13,272	2,220	60
total		20,585	3,220	100

Table 1

Conclusions

Students grade-level tests have unacceptable measurement error, especially with low-achieving students. The functioning-level test measures are two to three times more accurate than the grade-level test scores for predicting low-achieving students' achievement. This raises concern over the continued use of grade-level tests for student placement and school program evaluation.

When used with the same students, functioning-level tests like those used in the Portland Schools give more accurate assessments than the grade-level tests.

Functioning-level tests using item banks in which all items are calibrated to a single scale of difficulty accurately test students from the lowest grade-three level to the highest grade-eight level, Figure 3.

Figures 1, 2, and 3 are on next page (68).

George S. Ingebo, Ph.D.

Dr. Ingebo brings a wealth of life experience to student achievement testing. He grew up in Winnett, Montana, a small town in Petroleum County, Montana, where he learned to box at Shorty's Gym. During WWII, he flew combat missions with a B-24 bomber crew in India and China. He earned a Ph.D. from the University of Washington-Seattle. He taught high school science and mathematics, coached football and track.

A pioneer in constructing standardized machine scored tests, Dr. Ingebo established a college entrance testing program and developed predictors for college success. He directed a child clinical testing service at the University of Pacific. In 1969 Dr. Ingebo switched directions in testing. After hearing a lecture on Rasch Model testing by Ben Wright, Dr. Ingebo introduced this model into the Portland Schools. He established a new school testing program in Portland Elementary Schools. He provided technical planning in the Portland Metropolitan Area School Districts' High school testing cooperative. He helped found the Metropolitan Districts' Northwest Evaluation Association. He developed a variety of techniques for program evaluations based on Rasch equal interval measures from levels tests. He conducted research on the use of the Rasch Model over a 16-year period. Following on the Portland success, the Rasch model is increasingly used for measuring student achievement in metropolitan school districts.

Standard Deviation of Differences Between Expected and Actual Student Performance

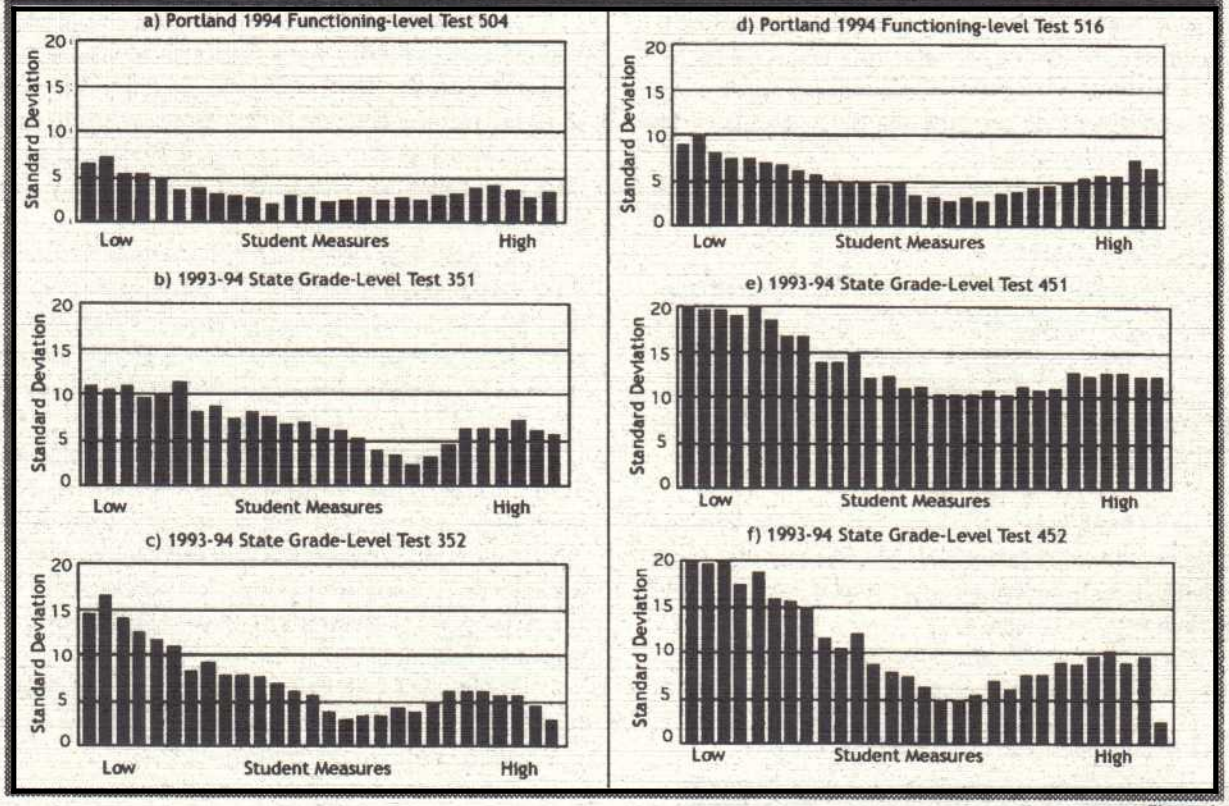
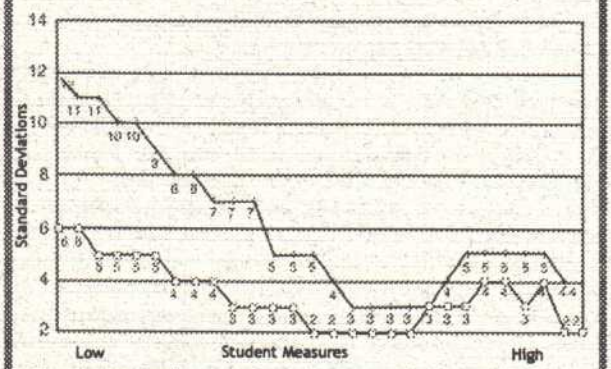


Figure 1

Plots of Aggregated Standard Deviation Differences of Expected and Actual Performance for All Students Taking 1993-94 Grade-Level and 1994 Functioning-Level Tests

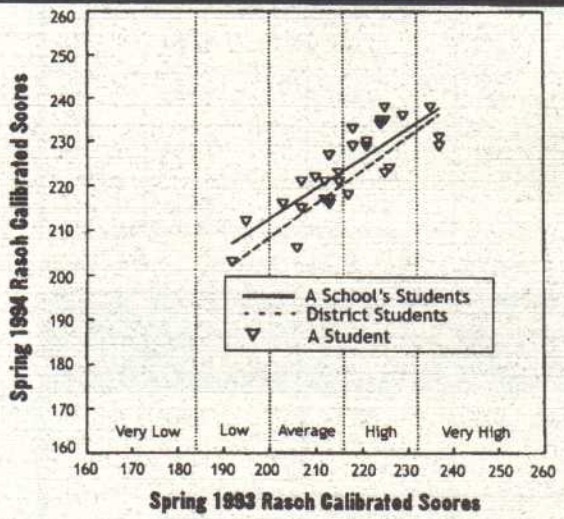
Reading and Mathematics Groups Combined



Functioning-level test accuracy improves with the middle and upper level students.

Figure 2

1994 Portland Public Schools Student Gains Compared to 1993 Test Scores



Comparing student triangles to the reference lines shows whether their gain is greater (above the line) or less (below the line). Comparing the two reference lines shows whether, on the average, students in a school gain more or less than similar level students in the rest of the District

Figure 3

Biological Evolution: A Tough Nut for Biology Teachers in Singapore?

Yew-Jin, Lee and Oon-Chye, Yeoh

Why the Nut is Important

If there can be one basic theme which gives Biology a unifying coherence, it has to be the theory of evolution. Evolution provides a philosophical system that aids our understanding of the living world; by experiment and theory. The famous geneticist Dobzhansky put it well when he said that nothing made sense except in the light of evolution (Dobzhansky, 1973).

Contrary to expectation, research into teaching and learning about this major concept in biology has been sparse despite its centrality to science literacy (Cummins, Demastes and Hafner, 1994). Indeed, the little that has been conducted has largely been in exploring students' ideas (which usually are quite inadequate in understanding) or on teaching for conceptual change in the sub-areas of evolution, for example: natural selection, competition, and population dynamics (see Demastes, Settlage and Good, 1995).

The first author is a practicing high school biology teacher, and it has been his experience to encounter much difficulty in getting his message across to his students in this topic. One key factor, besides the prevalence of students' alternative conceptions, was his less than satisfactory appreciation of this complex topic. However, he was not alone, as the lack of knowledge in evolution among teachers has been widely reported in the literature by Cummins, Demastes and Hafner (1994).

We use the metaphor on the understanding of evolution as a tough nut to crack for teachers. We are suggesting that teachers themselves are often unsure of evolution and that this lack of competency might be a hindrance in student learning. The many elegant concepts associated with evolution are truly 'choice morsels' of knowledge and understanding, but remain poorly taught in classroom learning and teaching as they were often beyond the grasp of teachers. The nut has not yielded its substance for many teachers!

Therefore, we wanted to find the levels of knowledge in local high school teachers (senior and junior high) regarding biological evolution and its subconcept of ecology. If their levels of comprehension were low like other teachers around the world, it would be understandable that students would also find this topic to be difficult to learn. Finding this basic but important piece of information would have implications for teacher education and reeducation.

We Don't Need No Nut Crackers!

A mailed survey questionnaire to 70 teachers was developed. It consisted of 36 five-option multiple-choice questions (MCQ) on evolution and ecology, while other sections gathered some demographic data. The MCQ were based largely on assessment questions from the Cambridge General Certificate in Education (GCE) 'Ordinary' (at grade 10) and 'Advanced' (at grade 12) level biology examinations that had a wide area of content coverage over these different levels of cognition. Twelve items were on ecology, with 25 on evolution. Both types of questions were randomly mixed in the first part of the questionnaire. The MCQ section was unspeeded in order to discourage guessing behavior and was intended to be completed in one continuous session by the high school biology teacher.

As expected in the mailed survey questionnaire, the responses took a long time to be returned, in fact five months! Twelve senior high (SH) and 40 junior high (JH) teachers completed the forms. This was 75% of the intended sample of 70 biology teachers in the study. We used QUEST version 2.0 computer program (Adams and Khoo, 1996) to apply the Rasch model to our data. This obtained an objective representation of the ability levels of the two groups of teachers placed alongside the difficulty levels of the items on the same scale in logits.

Who Had Cracked the Nut?

Rasch analysis (see **Figure 1** on next page) of items (M=0.0 logits, SD=1.23 logits) showed that items had a wide spread in terms of difficulty from 3.05 to -3.10 logits. There were no items misfitting the Rasch model except for Question 1 which QUEST could not calculate, as it had obtained a perfect scoring.

1) ***How much did the teachers know about biological evolution?***

The average raw score for the entire sample over all 36 items was 25.0 (69.6%) with a standard deviation of 4.9 ; lowest raw score was 15 and the highest 34. Generally, the teachers had performed better than average. The average score for SH teachers was 79.4%, and JH teachers was 66.6%. The range of logits for SH teachers was from 3.53 to 0.33, and for JH teachers from 3.01 to -0.56. The variable map in Figure 1 shows that these teachers' ability levels were higher than the difficulty of many of the items.

VARIABLE MAP

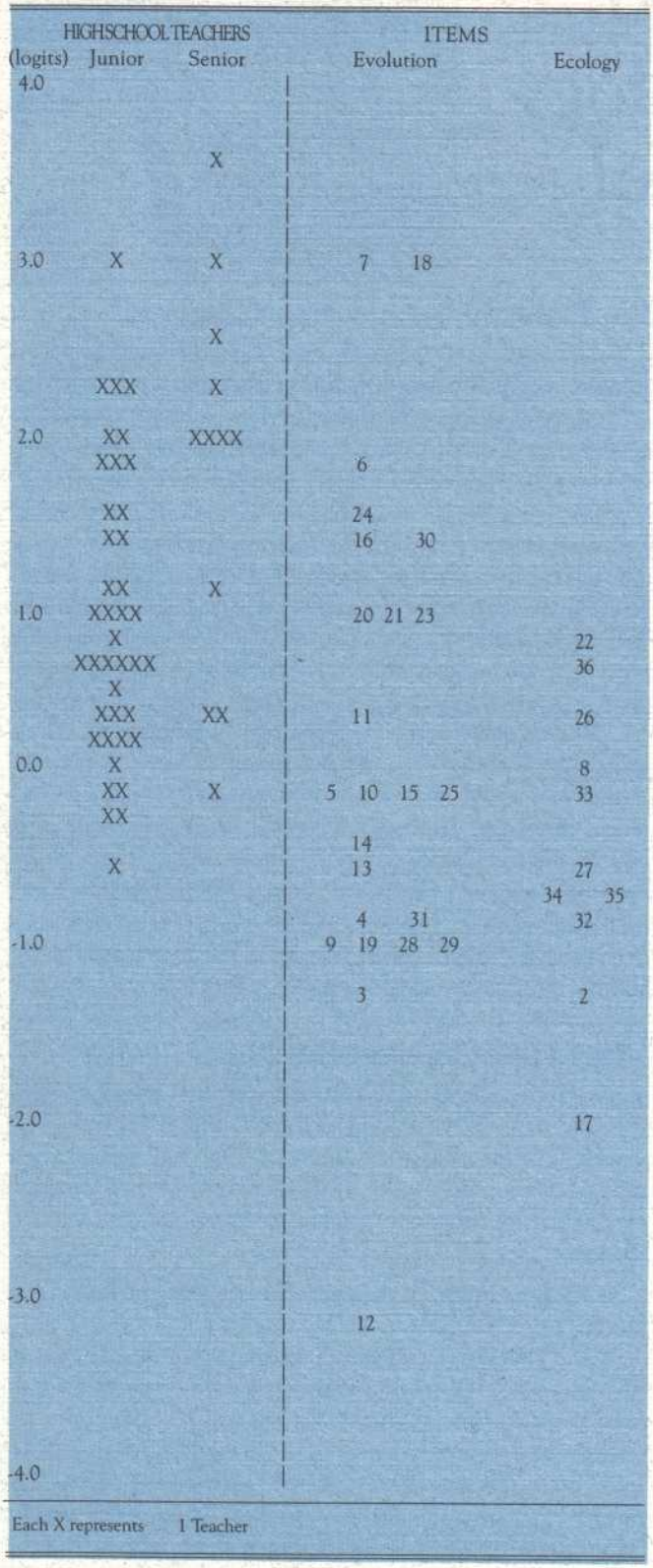


Figure 1. Item/Case map based on the responses by the junior and senior high biology teachers.

2) ***Do senior high or junior high teachers have better understanding?***

A t-test to compare the mean performances for SH (M=79.4% or measure 1.86 logits, SD=1.0) and JH teachers (M=66.6% or measure 0.89 logits, SD=0.9) was significant at $p<0.005$ level in favor of SH teachers. Figure 1 shows the item/case map of the two subpopulations of teachers, together with the items on evolution and ecology on the logit scale. Similarly, SH teacher raw scores over the sub-section of evolution **(Table 1)** were also significantly higher ($p<0.001$) but not with regard to ecology. The evidence suggests that SH biology teachers have a better understanding than JH teachers with regard to evolution, as was to be expected.

There were eight misfitting cases in the Rasch model (teacher nos. 6, 19, 20, 38, 42, 44, 50, and 51) according to **Figure 2** INFIT MNSQ (Adams and Khoo, 1996). However, none of the case values were more than 0.42 beyond the expected INFIT MNSQ value of 1.0, and only three exceeded 1.3.

QUEST produces KIDMAPs of cases. These are graphical representations of response patterns of the respondent together with the ability levels and the item difficulties on the same scale. These maps can help detect guessing behavior in the respondents.

A teacher who showed possible guessing behavior was case 42. According to **Figure 2**, the response behavior of this teacher was rather different from the rest of the teachers because they had a higher than expected INFIT MNSQ.

Examination of case 42's KIDMAP in **Figure 3** reveals many items in both Harder Not Achieved and Easier Not Achieved categories. This might indicate guessing behavior. Notice that while lucky guessing may have gained three right answers, unlucky guessing lost five improbable wrong answers. These features make KIDMAPs a valuable feedback mechanism for respondents in any test.

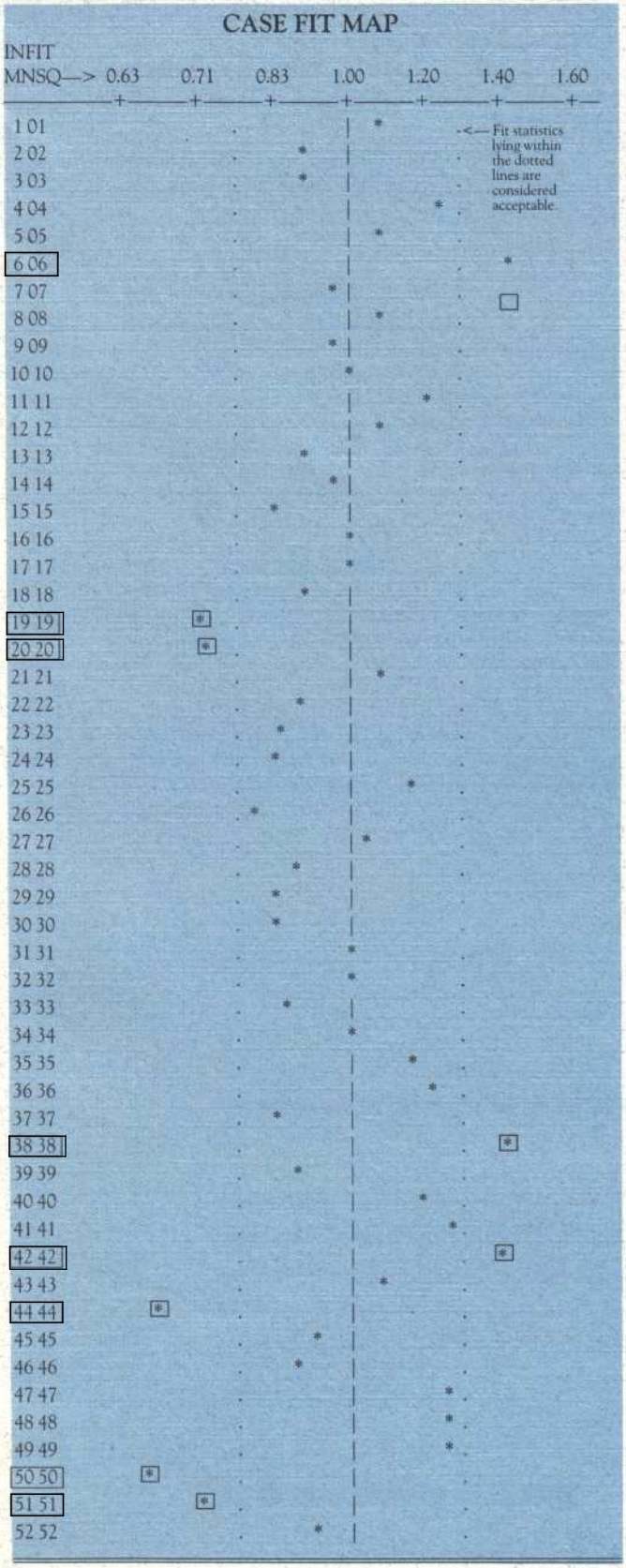


Figure 2. QUEST data output of the Case Fit Map.

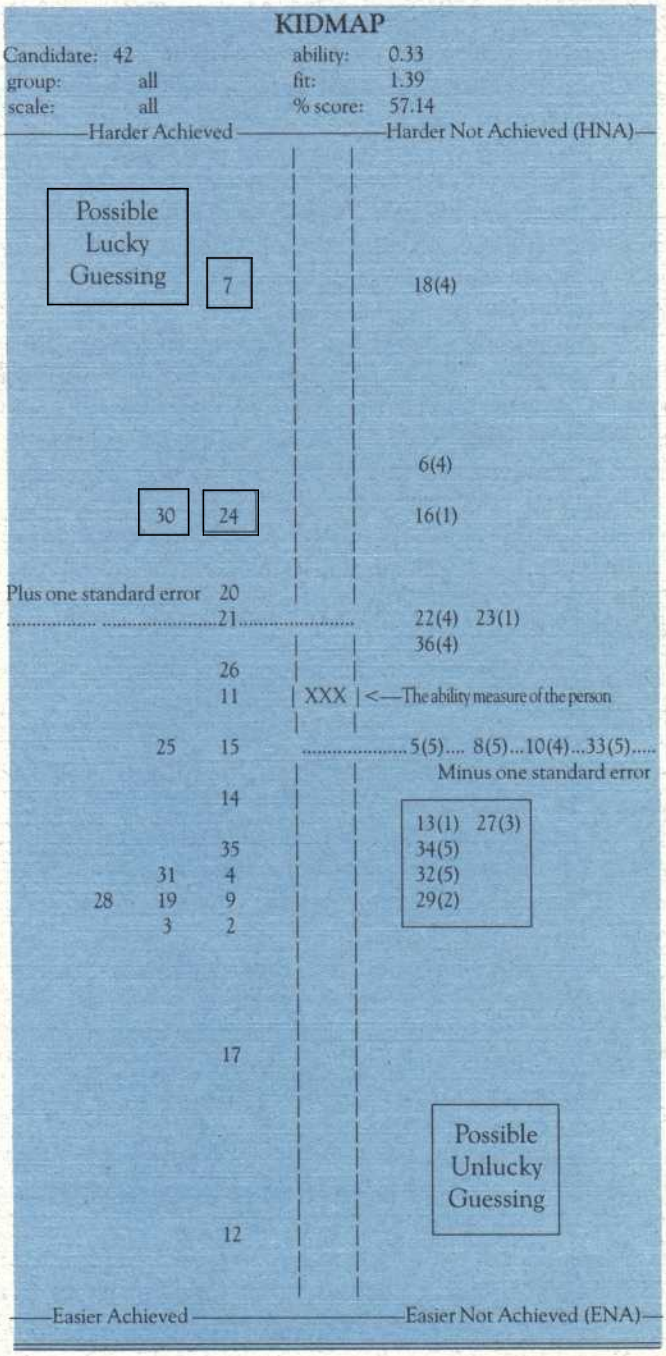


Figure 3. KIDMAP of Case number 42 who exhibited probable guessing behavior in answering the achievement test.

Conclusions

These results show that the graduate biology teachers (holding the General and Honors degrees) in Singapore had a reasonably good grasp of content in evolution and ecology. The ability measures from 52 teachers ranged from 3.53 to -0.56 logits (M= 1.12 logits, SD= 0.97); that of senior high teachers from 3.53 to 0.33 (M=79.4% or 1.86 logits, SD=1.0), and for junior high teachers from 3.01 to -0.56 (M=66.6% or 0.89 logits, SD=0.9). Senior high teachers had a significantly bet-

ter grasp of content than junior high teachers over the test as a whole ($p<0.005$) and over the subsection on evolution ($p<0.001$). There was no significant difference with regard to the section on ecology.

We recommend the Rasch model as a simple but valuable tool in the teacher's everyday repertoire of test data analysis. In this example Rasch allowed the objective determination of the levels of comprehension of evolution among teachers. We do not see any difficulty in using this method to elucidate ability levels of students as well.

Among other things, Rasch analysis can perform scoring, simple descriptive data analysis, and detect case and item 'misfits.' Though the relatively high levels of knowledge in local biology teachers were a cause for satisfaction, we discovered a number of alternative ideas or misconceptions, especially among junior high teachers. The QUEST KIDMAPs furnish valuable feedback to respondents in terms of showing their ability levels with respect to item difficulties, and can help detect the individual's pattern of answering and misconceptions.

References

Adams, R. J. and Khoo, S. T. (1996). Quest- the interactive test analysis system. Australian Council of Educational Research.

Cummins, L.C., Demastes, S.S. and Hafner, M.S. (1994). Evolution: biological evolution's under-researched unifying theme. Journal of Research in Science Teaching, 31(5), 445-448.

Demastes, S.S., Settlage, Jr., J., and Good, R. (1995). Students' conceptions of natural selection and its role in evolution: Cases of replication and comparison. Journal of Research in Science Teaching, 32(5), 535-550.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. American Biology Teacher, 35, 127-129.

Note: Part of the findings were from a Master's thesis in education submitted in December 1997 by the first author. It was titled 'Comprehension of Biological Evolution by High School Biology Teachers in Singapore.

Yew-Jin, Lee

Yew-Jin, Lee has a background in Zoology; his final year university dissertation was on local sea slug taxonomy. Among his current interests are collecting invertebrate and fish fossils; mucking around with computers and the Internet; and soothing the mind with World and Indie music. For a living, he is a high school biology teacher in Singapore and is also in charge of student welfare and counseling. He found time to combine his interest in biology with measurement theory in a Masters degree in education, of which part of the results of his thesis are published in the article. Victoria School, 3 Geylang Bahru Lane, Singapore 339626.

Dr Oon-Chye, Yeoh

Oon-Chye,Yeoh is a teacher educator at the Nanyang Technological University, Singapore. As the science educator (Ph.D Stanford U,1973), his research and teaching interests embrace Curriculum and Instructional Design, Curriculum Evaluation, the Psycho-pedagogy of Teaching and Learning, Environmental Education and, in particular, the Implementation of Curriculum Change. Over 40 years, he's offered consultancies in many of the East and S.E Asian countries under UNESCO, UNICEF, and SEAMEO oranizations. He was the project director and, later, the member of Singapore's Second (1984-89) and the Third (1994-96) International Maths and Science Studies, respectively. After 40 years, he still finds teaching at all levels a major challenge, but at the same time most satisfying while learning from his lively and brilliant students. School of Science, National Institute of Education, 469 Bukit Timah Road,Singapore 259756.

Hertz, the "German physicist who first verified James Clerk Maxwell's electro-magnetic equations by demonstrating the existence of radio waves" said that "One cannot escape the feeling that these mathematical formulae have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverers, that we get more out of them than was originally put into them!" (Dyson 1969: 99)

A Second Scoring Mechanism to Study Change

Winifred A. Lopez, Ph.D.
International Survey Research LLC

Fullan (1991) reminds us that "change is multi-dimensional" and that it can "vary within the same person as well as within groups." Yet research frequently conceptualizes change solely in terms of differences in status. Researchers gather data over two time points, compute the status for each time point, and proceed to look for differences in status over time.

Consider the example of a survey of 7 items designed to measure teacher dynamism in 89 urban elementary schools. The survey administered in 1991 and 1994 consisted of items that probed aspects of teacher behavior ranging from the more passive behaviors like acceptance of instructional goals and familiarity with school improvement plans, to more active behaviors like voicing concerns, enforcing rules, and helping to develop school improvement plans. Teachers responded to these items using a four-point rating scale, "Strongly Disagree, Disagree, Agree, Strongly Agree." These responses were scored 1,2,3, and 4 respectively. To study how

Figure 1: Teacher Dynamism - Patterns of Change 1991-1994

Acquiescence : Average94 - Average91

Teacher Dynamism Average94 - Average91	Decrease (-)	No Change (0)	Increase (+)
Increase (+)	Moderates Progress	(1 school) All Round progress	(4 schools) Laggards Progress
No Change (0)	Radicalizing	(58 schools) Static	(21 schools) Homogenizing
Decrease (-)	Moderates Deteriorate	(3 schools) All Round Deterioration	(2 schools) Progressives Deteriorate

Note:
1=Strongly Disagree, 2=Disagree, 3=Agree, 4=Strongly Agree
Laggards= Those who pick Categories 1 & 2 (Strongly Disagree, Disagree)
Moderates= Those who pick Category 3 (Agree)
Progressives= Those who pick Category 4 (Strongly Agree)
Arrows represent movement from one category into another
The topmost entry in each cell represents the number of schools

teacher dynamism in schools had changed from 1991 to 1994 we used BIGSTEPS (Wright & Linacre, 1995) to assign a measure of dynamism to each teacher, aggregated these measures to the school level, and computed differences in average measures over time. T-tests at the 5% significance level showed that there was a significant increase in teacher dynamism in 5 schools (6%), a significant decrease in another 5 schools (6%), and no significant change in 79 (88%) schools.

While interesting, the analysis is limited in that it focuses only on the magnitude and direction of change. We learn nothing about the patterns of change. For instance, the condition of no change can occur in one of two ways. "It would result if everyone is doing about the same as before" or if "gains in some parts are being canceled out by losses elsewhere" (Namboori et al., 1993). Similarly, we do not know what patterns underlie schools where teacher dynamism increased. Specifically, we do not know whether improvement was universal or if it was restricted to certain groups of people.

The discussion that follows outlines how two scoring mechanisms can be used in conjunction with each other to study patterns of change. This follows from Guttman (1950) who proposed that a construct be investigated from two points of view, "a content scale which ranks people from high to low on a single continuum" and "an intensity scale which ranks people from high to low on a single intensity continuum." The technique used BIGSTEPS (Wright & Linacre, 1995) and the following scoring mechanisms to create two measures for each person from the same set of items, a measure of teacher dynamism and a measure of acquiescence.

	Strongly Disagree	Disagree	Agree	Strongly Agree
Teacher Dynamism Scoring	1	2	3	4
Acquiescence Scoring	0	0	1	0

School means were computed for teacher dynamism and acquiescence for 1991 and 1994, and T-Tests performed for both aggregate measures. This allowed us to classify each school into one of three categories (significant increase, significant decrease, and no change) for teacher dynamism and acquiescence. By considering changes in acquiescence alongside changes in teacher dynamism, we can identify 9 different patterns of change (Figure 1). For example, a decrease in acquiescence from 1991 to 1994 suggests movement out of the "Agree" category. If the decrease in acquiescence occurs alongside an increase in teacher dynamism, we can deduce that the net movement must be from a less favorable category into a more favorable one (cell labeled "Moderates Progress"). If teacher dynamism remains the same while acquiescence decreases, we can deduce that the movement out of the "Agree" category is bidirectional (Cell labeled "Radicalizing"). The level of teacher dynamism remains unchanged because gains in one direction (movement from Agree to Strongly Agree) are offset by losses in the other (movement from Agree to Disagree and Strongly Disagree).

We are now able to expand our original classification of schools. Of the 79 schools classified as showing no change in teacher dynamism levels, 58 can be considered static with respect to category use patterns. These schools did not show a net movement either into or out of the "Agree" category. In the remaining 21 schools there was a homogenizing trend, i.e., a net movement into the "Agree" category from the more positive "Strongly Agree" category as well as from the negative "Strongly Disagree" and "Disagree" categories. Two patterns of change occurred in the 5 schools that showed a decrease in levels of teacher dynamism. In 3 of these schools there appeared to be an all-around deterioration. Two schools showed localized deterioration, characterized by a movement out of the "Strongly Agree" category to the "Agree" category. Finally, among schools which showed an increase in teacher dynamism, one showed all-around improvement, while 4 schools showed restricted improvement, characterized by a movement out of the "Strongly Disagree" and "Disagree" categories into the "Agree" category.

References:

Fullan, M. (1991). The new meaning of educational change. Columbia University: Teachers College.

Guttman, L. (1950). The intensity component in attitude and opinion research. In S.A. Stouffer et al. (Eds.). Measurement and prediction (pp 213-276). New York: Wiley.

Namburi, K., Cortin, R. & Eberst-Dorsten (1993). Analyzing distributions in school effects research: An empirical illustration. Sociology of Education, 4, 278-294.

"What the human sciences require for more dramatic progress [is] not simply more data (of the same kind), as so many empiricists have stated, but new instrumentation for obtaining data, or reasonable theoretical restrictions of data domain so that more exhaustive explanatory possibilities can be tried." Ackermann, John R. Data, instruments, and theory: a dialectical approach to understanding science. Princeton, New Jersey: Princeton University Press, 1985, p. 169.

TEACHING

PAY ATTENTION!!!

Screening for Attention Deficit Hyperactivity Disorder in College Students

Everett V. Smith, Jr., Ph.D.
The University of Connecticut

"Imagine living in a fast-moving kaléidoscope, where sounds, images, and thoughts are constantly shifting. Feeling easily bored, yet helpless to keep your mind on tasks you need to complete. Distracted by unimportant sights and sounds, your mind drives you from one thought or activity to the next. Perhaps you are so wrapped up in a collage of thoughts and images that you don't notice when someone speaks to you" (Sharyn, 1994).

This description is what 3 – 5% of all children feel like, with approximately three times the number of boys being affected. These children are dealing with Attention Deficit Hyperactivity Disorder (ADHD). ADHD is a disorder with no physical signs. It can only be identified by looking for certain behaviors. These behaviors are characterized by inattentiveness (e.g., failing to complete assignments), impulsivity (e.g., interrupting conversations), and hyperactivity (e.g., always in motion, restless, fidgeting). The diagnostic subtypes are predominantly inattentive, predominantly hyperactive/impulsive, and combined inattentive and hyperactive/impulsive.

For years ADHD was considered a childhood diagnosis that was outgrown. However, in recent years many late adolescents and adults have sought help for ADHD. It has been estimated that 40-80% of children with ADHD continue to experience symptoms into late adolescence and adulthood. The Diagnostic and statistical manual of mental disorders (DSM-IV: APA, 1994) criteria requires the same number of symptoms for the diagnosis regardless of age. However, college students (the focus of this research) were not even included in the DSM-IV field trials. Given the increasing number of college students seeking ADHD evaluations (at one large southwestern university the number of students receiving services for ADHD has increased 150% in the past 3 years and the number requesting evaluations for ADHD has increased 300%) and mandates for colleges and universities to provide services, studies providing evidence for the validity of the DSM-IV criteria with college students are needed.

Smith and Johnson (in press) investigated the dimensionality of the 18 DSM-IV symptoms in a college sample using The Adult Behavior Checklist (ABC: Johnson & Lyonfields, 1995). The ABC is a self-report assessment that is designed to screen for ADHD as defined by the DSM-IV criteria. The symptoms were reworded in order to allow participants to rate the overall frequency of their behavior. Results of the Smith and Johnson investigation led the researchers to conclude that 15 of the 18 symptoms could be used to reflect the hypothesized dimensions of inattention and hyperactivity/impulsivity.

Should you be evaluated for ADHD?

Using the inattentive items from the ABC, the following is a demonstration of how a qualified individual may use the ABC to decide whether or not you should receive a more extensive evaluation. For each item in Table 1, begin by asking yourself "During the past six months ...". If you feel that the behavior NEVER occurs, give yourself a '1' for that item, if you feel the behavior SOMETIMES occurs, give yourself a '2', if the behavior OFTEN occurs a '3', and if the behavior occurs VERY OFTEN, a '4'. Repeat this process for all nine items addressing inattentiveness and add the resulting values. Using Table 2, which is a raw score to measure conversion table provided courtesy of BIGSTEPS (Linacre & Wright, 1995), find the value of your raw score under the column labeled 'Score.' Find the corresponding linear measure under the column labeled 'Measure'. This value represents the estimated amount of the latent trait 'inattentiveness' which you possess. Is this value high enough to warrant a full evaluation? The current cutoff for further evaluation corresponds to a measure of .39 logits (i.e., the measure corresponding to a raw score of 25 minus one standard error). If your estimated measure is greater than or equal to .39 logits, you may be a candidate for further evaluation. If your measure is below .39 logits, add the standard error corresponding to your estimated measure (found in the column labeled 'S.E.') to your estimated measure. This attempts to account for chance fluctuations in the estimation of your 'inattentiveness' measure. If this value is greater than or equal to .39, you may be referred for a more comprehensive evaluation.

Table 1

Item content of the Adult Behavior Checklist

Item 1 -	You fail to pay close attention to details or make careless mistakes in school, at work, etc.
Item 2 -	You have difficulty sustaining your attention to tasks or in play activities.
Item 3 -	You do not listen when directly spoken to.
Item 4 -	You do not follow through on instructions and fail to finish schoolwork, chores, work duties, etc.
Item 5 -	You have difficulty organizing tasks and activities.
Item 6 -	You avoid, dislike, or are reluctant to engage in tasks that require sustained mental effort (e.g., homework or schoolwork).
Item 7 -	You lose things necessary for tasks or activities (e.g., books, school assignments, tools or keys).
Item 8 -	You are easily distracted by extraneous stimuli (e.g., traffic noises, conversations, or looking out the window).
Item 9 -	You are forgetful in daily activities.
Item 10 -	You have difficulty playing or engaging in leisure activities quietly.
Item 11 -	You are "on the go" or act as if "driven by a motor".
Item 12 -	You talk excessively.
Item 13 -	You blurt out answers before questions have been completed.
Item 14 -	You have difficulty awaiting your turn.
Item 15 -	You interrupt or intrude on others (e.g., butt into conversations or activities).
Note:	Items 1 through 7 are for the inattentive dimension; items 8 through 15 are for the hyperactive/impulsive dimension.

of other conditions (e.g., depression, anxiety, substance abuse) is also necessary.

Information has been provided for the inattentive set of items from the ABC. Comparable information for the hyperactivity/impulsivity dimension was not provided, as the intent of this article is not to promote self-diagnoses of ADHD, but rather to demonstrate some of the potential benefits of Rasch measurement. Further benefits of Rasch measurement would be realized when gauging the progress (i.e., comparison of pre/post measures) of therapy. In this situation, Rasch measurement would provide the interval level measures necessary for arithmetic operations and many statistical methods. The method presented in this article fails to directly take into account the influence of outlying and extreme responses on the measurement process. See Linacre (1997) for the use of expected score maps to handle these situations. Interested readers and researchers may contact the author for additional information on the ABC and its current state of development. A comprehensive on-line source is also available from the National Institute of Mental Health at http://www.nimh.nih.gov/publicat/adhd.htm.

Table 2

Raw score to linear measure conversion table

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
9	-5.47E	1.47	19	-.62	.55	29	1.69	.47
10	-4.69	1.08	20	-.33	.53	30	1.92	.49
11	-3.82	.82	21	-.07	.51	31	2.17	.51
12	-3.24	.72	22	.18	.49	32	2.45	.55
13	-2.77	.66	23	.42	.48	33	2.78	.61
14	-2.35	.63	24	.64	.47	34	3.23	.73
15	-1.97	.61	25	.85	.46	35	3.95	1.01
16	-1.61	.60	26	1.06	.46	36	4.66E	1.42
17	-1.26	.58	27	1.27	.46			
18	-.93	.57	28	1.48	.46			

Note: **Calibration based on 1503 participants.**

Discussion

Simple, inexpensive, and efficient screening assessments such as the ABC are useful for an initial diagnosis of ADHD. However, only a comprehensive ADHD evaluation done by a trained professional (e.g., a psychiatrists, psychologists, pediatricians, or neurologists) can yield more definitive evidence of an ADHD diagnosis. This type of evaluation may include a review of medical, family, and academic records as well as formal assessments of intelligence, memory, and attention/concentration. In addition, the behavioral symptoms need to be confirmed with someone familiar with the individual's behavior (e.g., spouse, parents, roommates). Careful consideration

References

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, D.C.: Author.

Johnson, B.D., & Lyonfields, S. (1995). ADHD and college students: How useful are the DSM-IV criteria? Paper presented at the 103rd annual Convention of the American Psychological Association, New York, NY.

Linacre, J.M. (1997). Instantaneous measurement and diagnosis. In R.M. Smith (Ed.), Physical Medicine and Rehabilitation: State of the Art Reviews (pp. 315-324). Philadelphia: Hanley & Belfus, Inc.

Linacre, J.M., & Wright, B.D. (1995). BIGSTEPS computer

PSYCHOLOGY

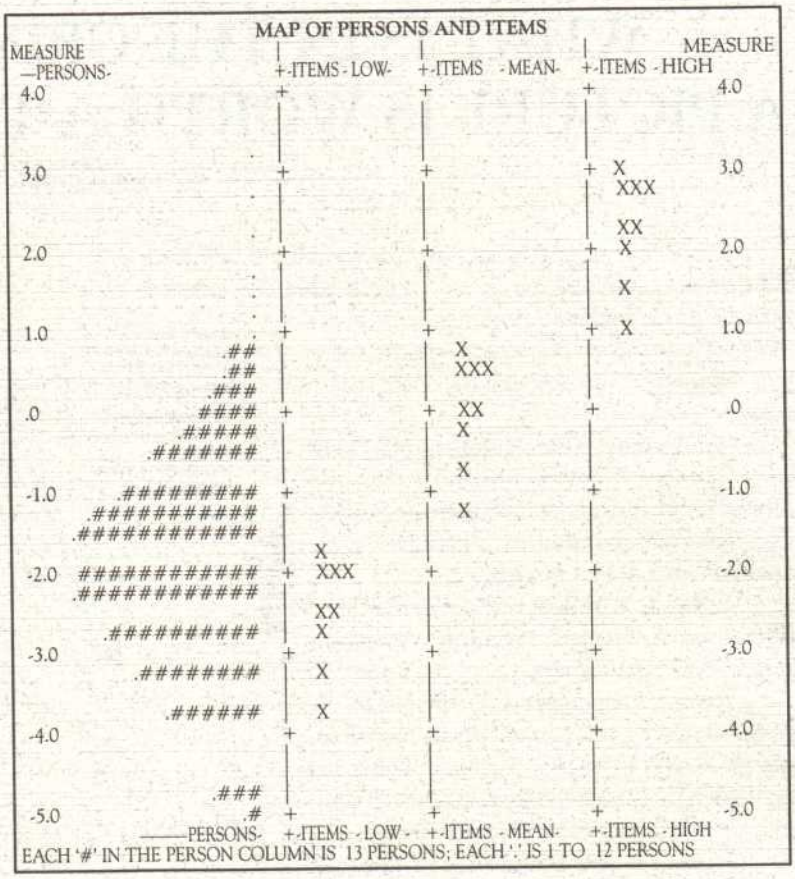
program. Chicago: MESA Press.

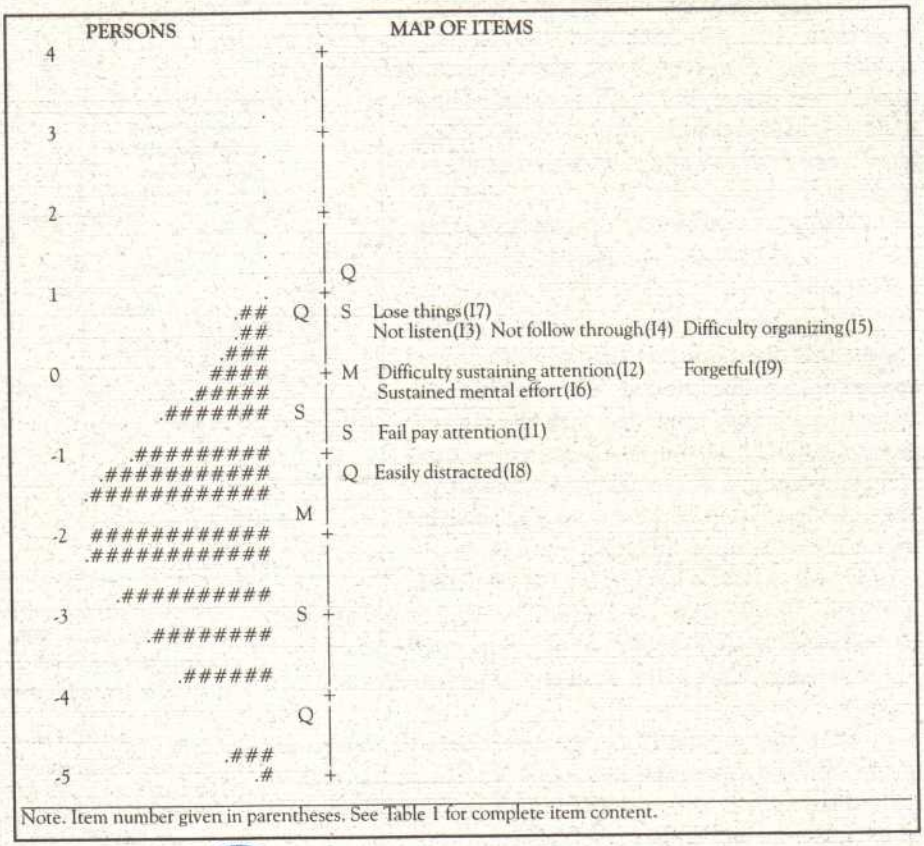
Sharyn, N. (1994). Attention Deficit Hyperactivity Disorder. Available online: http://www.nimh.nih.gov/publicat/adhd.htm.

Smith, Jr., E.V., & Johnson, B.D. (in press). Factor Structure of the DSM-IV Criteria for College Students using the Adult Behavior Checklist. Measurement and Evaluation in Counseling and Development.

Author Note

Correspondence concerning this article should be addressed to Everett Smith, University of Connecticut, Department of Educational Psychology, U-64, 249 Glenbrook Road, Storrs, Connecticut, 06269. Electronic correspondence may be addressed to esmith@uconnvm.uconn.edu.





Everett V. Smith, Jr.

Education:

B.S. Applied Mathematics, 1991, The University of Connecticut

M.A. Teaching, 1992, Sacred Heart University

Ph.D. Educational Psychology, 1995, The University of Connecticut

Recent Positions:

Lecturer, Educational Psychology, The University of Connecticut, 1997 – present

Psychometric Consultant, Hartford Hospital, 1997 – present

Assistant Professor, Educational Psychology, The University of Oklahoma, 1995 - 1997

Courses Taught:

Principles of Measurement and Evaluation, Quantitative Research Methods I and II, Applied Multivariate Statistics, Research Methodology, and Affective Instrument Design

Research Interest:

Applications of objective measurement in cognitive and counseling psychology. Specific interest in the measurement and application of social cognitive constructs.

Other interest:

Tennis, mountain biking, and spending time with my golden retrievers, Calvin and Hobbes (and yes, they unfortunately behave like the comic strip characters).

WHAT IS IN THE CRIMINAL'S MIND? A PICTURE IS WORTH A THOUSAND WORDS

By George Karabatsos

George Karabatsos, MESA Psychometric Laboratory

Measurement is very useful when meaning is provided. For instance, in describing rapists' attitude towards women, it is vague if one concludes with a numerical answer, such as the average questionnaire score. It is more helpful to describe the specific attitudes he possesses.

To illustrate, a Rasch analysis of a "Hostility Towards Women" inventory was performed for Rapists and Non-Rapists separately. The item calibrations of each group are compared on an X-Y Plot in Figure 1.

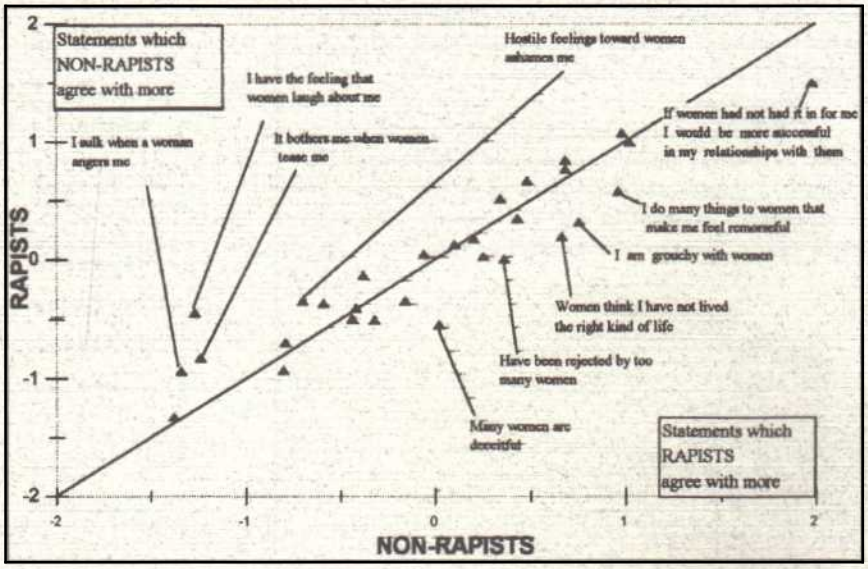


Figure 1. Item-identity plot of the "Hostility Towards Women" inventory, comparing Rapists and Non-Rapists.

Items in the upper left triangle of the graph represent statements which Non-Rapists agree with more than Rapists. The lower right triangle contains statements which Rapists agree with more than Non-Rapists.

Figure 1 is very informative. Rapists appear to feel criticized, abandoned, and rejected by women. Also, they distrust and hold animosity towards the opposite sex. Non-Rapists are more likely to feel teased by women, and shame over hostile feelings. These findings provide evidence that Non-Rapists have more sense of shame and guilt than their counterparts.

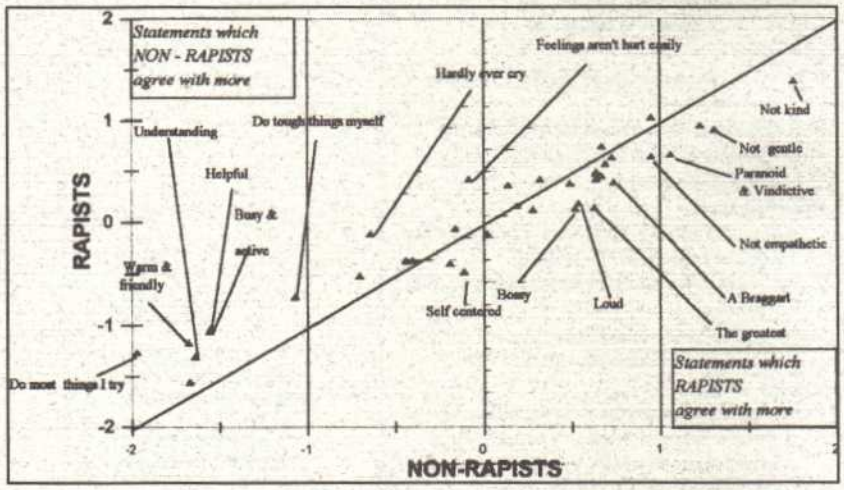


Figure 2. An item identity plot of the "Androgyny Scale" which compares Rapists and Non-Rapists.

What do rapists think of themselves? To answer this question, the same procedure was applied using a different questionnaire. Figure 2 shows that rapists, compared with the control group, are more paranoid, asocial, revengeful, insecure, immature, presocial, and narcissistic. These findings support the theory that sexually aggressive men have infantile traits of dominance, tend to be self-serving, are self-centered, vindictive, and not concerned with the welfare of other people. On the other hand, Non-Rapists are more confident, in control of their emotions, social, mature, and responsive to other's needs.

With two pictures alone, we have been able to learn so much. Although objective measurement is the cornerstone of good science, it is the meaning attached to measurements which generate and refine useful theories.

George is a Ph.D. student in the Measurement, Evaluation, Statistical Analysis (MESA) program at the University of Chicago. He is researching the connections between Rasch measurement and Axiomatic Measurement Theory, particularly the application of the Rasch model to measure non-additive representations.

His previous work with Rasch measurement involved the validation of a quality-of-life questionnaire for Multiple Sclerosis patients, research into the measurement properties of a sex crime survey, and the investigation of the factors that lead to sexual aggression and victimization.

e-mail: gkarabat@midway.uchicago.edu

PSYCHOLOGY