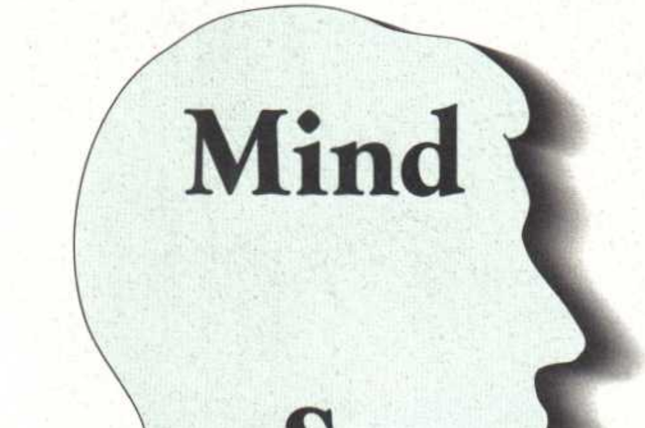
Popular Measurement

Volume 3 - No. 1 - $15.00
Spring 2000

Journal of the Institute for Objective Measurement



Inside:

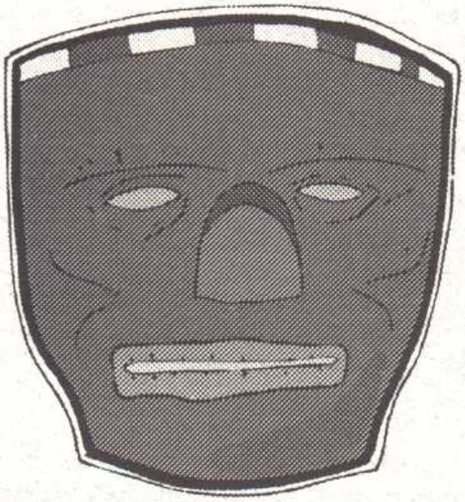
INSTITUTE FOR OBJECTIVE MEASUREMENT
155 N Harbor Drive #3104
Chicago, IL 60601

INDEX

5
14
27
34
40
53
58
70
72
78

Fechner: The Man in the Mask

Larry H. Ludlow and Rose Alvarez-Salvat
Boston College



Emaciated, nearly blind, alone by choice in rooms with blackened walls. Communicating through funnels doors while wearing a metal mask. Despondent and wishful of death yet persisting in volitional exercises to channel mental forces to subject his involuntary physical functions to voluntary control. Although dismissed by some as a mental patient, he gained renown as "the father of experimental psychology." So, who was this man in the mask?

Gustav Theodor Fechner is well known to us through "Fechner's Law." This law was one consequence of a lifelong interest in the potentialities of the mind, particularly in the relationship between the mind and body. This interest led him to argue that the mind (sensation) and body (stimulus) had to be regarded as two separate entities in order that each could be measured and the relation between the two determined (separation of parameters?).

He encountered a problem. While the magnitude of a stimulus can be directly measured, the magnitude of a sensation can not. But since we can physically measure the stimulus values that give rise to a sensation, we can indirectly measure sensation by taking differences between two stimuli. To determine the magnitude of a sensation, we then take the just noticeable difference (jnd) between two stimuli as the unit of sensation and count up jnd's from zero sensation at the absolute threshold to the sensation that is being measured. This reasoning led to: $S = k \log R$, where S (the magnitude of sensation) is the number of jnd that the sensation is above zero, R is the magnitude of the stimulus, and k is a proportionality constant. (Interestingly, the law requires the existence of negative sensations.)

With this formula, Fechner believed that the dualism between mind and body had disappeared and the nature of psychophysics as "an exact science of the functional relations or the relation of dependency between mind and body" had been established. For us, the significance of his work was that it took measurement beyond solely material phenomena to what Fechner referred to as the immaterial mind and the spiritual world.

So then, what drew Fechner to the study of the mind? Why, in particular, did he seem to have a fixation on measuring magnitudes of sensations? Was his interest purely academic and suited to the times (Zeitgeist) or

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

Rose M. Alvarez-Salvat

Rose M. Alvarez-Salvat is a doctoral candidate in the Counseling Psychology program at Boston College, Chestnut Hill, MA. Her research interests are in adolescent identity development, multicultural issues, and psychological assessment.

was there a personal interest in his quest? These questions prompted this article.

First, there were his spiritual, philosophical beliefs.

His father was the village pastor and his uncle, too, was a preacher. Both men contributed to his lifelong philosophical stance against materialism through their examples of independence of thought and receptivity of new ideas. His support of spiritualism as opposed to materialism is seen in his *The Little Book on Life after Death* (1836). His spiritualism even took him so far as to argue for the mental life of plants (Nanna, 1848).

Second, there were his medical studies.

At the age of 16, he went to Leipzig to study physiology, which meant studying for a doctorate in medicine. This was a time when medical thinking and practice were characterized by philosophical considerations and systems. There was an absence of an empirical basis in medical doctrine. It was a time of experimenting and waiting for chance hits. The practice of medicine was so chancy that Fechner adopted the pen name Dr. Mises and wrote numerous satires on current medical practices (seen in his *Proof that the Moon is Made of Iodine, 1821*).

His medical studies were, however, purposeful in another direction. He studied the brain and was intrigued by its duplex structure. He came to regard consciousness as an attribute of the cerebral hemispheres and he placed great stress on the equipotentiality of the cerebral cortex. In fact he argued that if it were possible to split the brain longitudinally it would achieve something like the duplication of a human being, in effect dividing the stream of consciousness.

Third, there was his professional career.

After medicine he studied physics and mathematics and was made professor of physics at Leipzig. During this period, he became acquainted with the work of Bernoulli and Laplace. From Bernoulli's probabilistic work linking fortune morale and fortune physique, he saw a mathematical relationship that corresponded exactly with his goal of connecting mind and body. From Laplace, he saw the value of applying the normal law of error in experimentation. Combined with these mathematical interests, he had a growing interest in sense-physiology, especially on complementary and subjective colors and subjective after-images. His experimental enthusiasm for gazing at the sun through colored glasses, however, permanently injured his eyesight. In 1839 he resigned his position due to poor health partly because of this injury.

Finally, there was his "life-crisis."

Fechner spent 1839 to 1851 in retirement. These were not pleasant years. He lost his health, his sight, his income, his friends, and even his wife at times. He suffered intense physical pain and mental anguish. He was profoundly despondent and obsessively brooding. He physically isolated himself and refused to eat for extended periods. But he would not give in to his suicidal wishes. His perspective was "If I put an end to my life here, I must make atonement and undergo all my sufferings in my future life". This attitude led to a system of experiments designed to mitigate his suffering and facilitate his healing.

First, he regarded the medical advice of the day as "fruitless" and began his own series of treatments. These consisted of stimulants, infusions, draining remedies, electrical applications, steam baths, opium, and even animal magnetism. These were all without success.

Second, he truly believed that mind and soul are the ultimate of reality, a philosophical position he called the "day view". Starting from this position allowed him to consider that the brain possessed powers not fully realized or explored. He was particularly concerned with establishing psychophysical functional connections that would guide him to personal psychotherapeutical procedures. For example, he suffered from an intense digestive disorder that transmitted sensation (pain) to his brain. He reasoned that if the digestive organs could transmit signals to the brain, "why not conversely, by the exercise of volition, bring about a conduction from the brain to those organs and thus remedy them?"

This reasoning led to a system of exercises to not only increase his mental effort at reducing pain but to turn back and heal the disorder in the first place. Here we see the stimulus (the volitional exercise of mental effort) and the sensation (the pain). What was their relation-was it one-to-one? How could he control the relation-could he simply concentrate harder and through auto-suggestion heal his condition?

On the morning of October 22, 1850 he had the insight that led to this date being honored as Fechner Day. While lying in bed puzzling over how to mathematically link body and mind (or stimulus and sensation, or mental effort and pain) he proposed that a geometric series in the intensity of a stimulus might correspond to an arithmetic series in the sensation. This idea (a direct consequence of his painful experiences?) established the program of research that Fechner called psychophysics.

Rather than succumbing to his condition and disappearing from history, he took advantage of his personal philosophy and professional training to extract meaning from his illness. His drive to model the world he lived in left us with methods of measurement we employ on a daily basis. In fact, the next time you use the remote on your television and adjust the red color back and forth until it's just right for you (but maybe not for your partner), you are applying the principles he established 150 years ago.

(To this day there is no definitive explanation of what his illness was nor how he was able to recover from it.)

References

Benjamin, L. T. (1997) (2nd ed.) A History of Psychology: Original Sources and Contemporary Research. New York, NY: McGraw-Hill.

Boring, E. G. (1950). (2nd ed.). Gustav Theodor Fechner. A History of Experimental Psychology (pp. 275- 296). Englewood Cliffs, NJ: Prentice-Hall.

Boring, E. G. (1961). Human Nature vs. Sensation: William James and the Psychology of the Present-1942. In Psychologist at Large: An Autobiography and Selected Essays (pp. 194- 233). New York, NY: Basic Books.

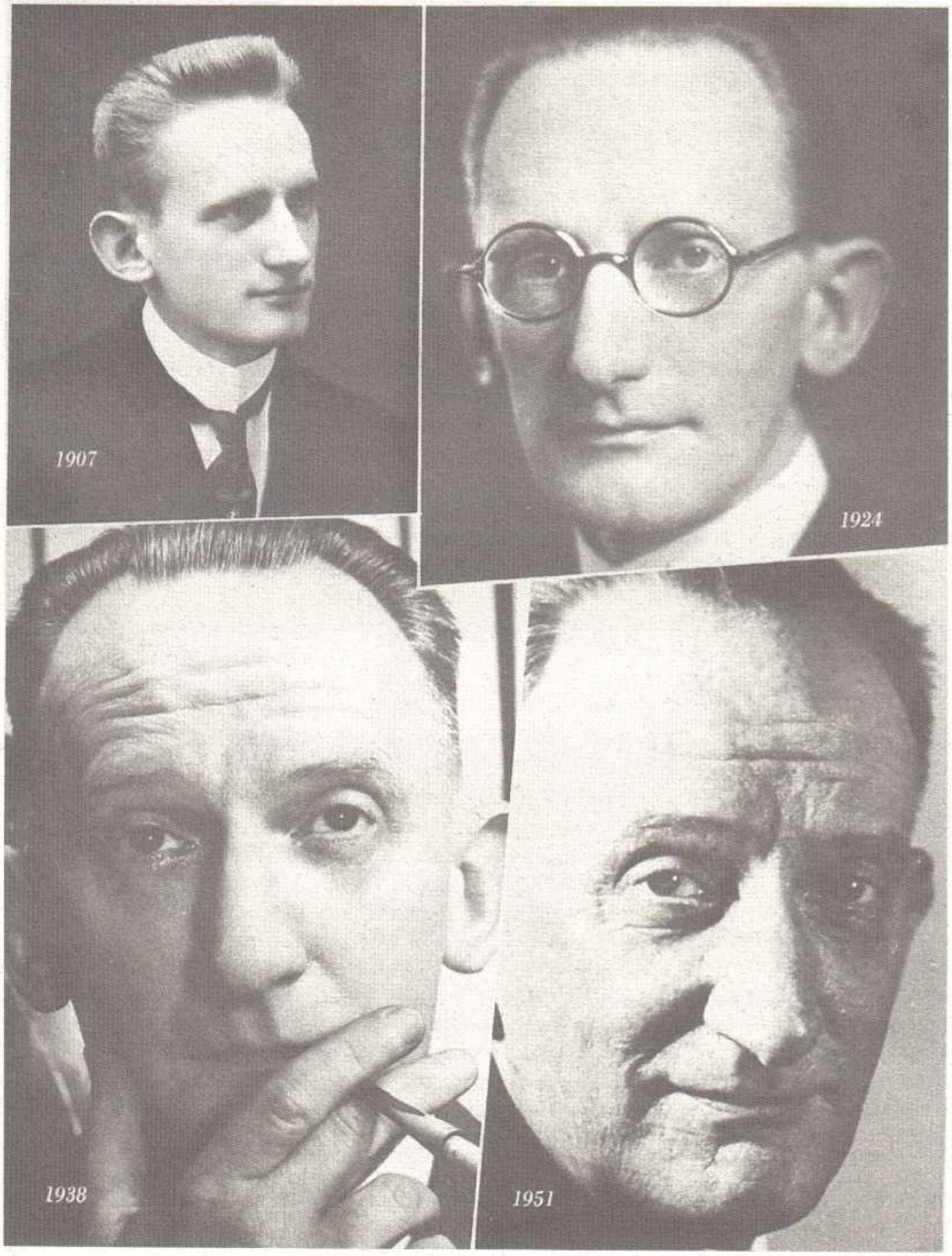
Boring, E. G. (1963). Fechner: Inadvertent founder of psychophysics. In R. I. Watson & D. T. Campbell (Eds.), History, Psychology and Science: Selected Papers (pp. 126-158). New York, NY: Wiley and Sons.

Schroder, H. & Schroder, C. (1993). Gustav Theodor Fechner during his life-crisis. In H. Schroder, K. Reschke, M. Johnson & S. Maes (Eds.), Health Psychology-Potential in Diversity. Regensburg: S. Roderer-Verlag.

Zangwill, O.L. (1976). Thought and the brain. British Journal of Psychology, 67, 301-314.

Thurstone:
Measurement For a New Science

Nikolaus Bezruczko, Ph.D.



"He stole fire from the gods, then paid with factor analysis."

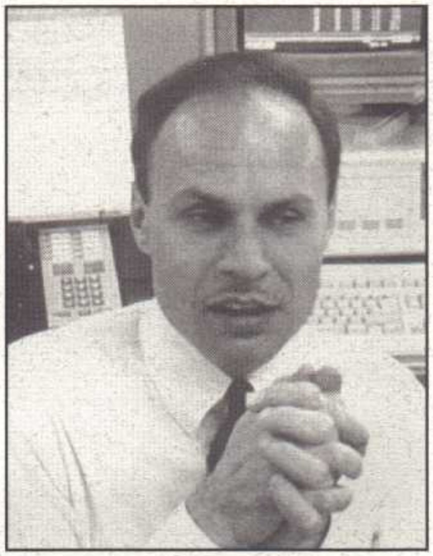
In ten short years, early in his career, Louis L. Thurstone revolutionized nonphysical scaling by single-handedly adapting the psychophysics developed by Fechner, Wundt, and Müller to measure mental forces in 20th century psychology. In contrast, the long, slow labor of factor analysis overwhelmed him for more than twenty years, as he tried to develop and defend it. His measurement advances were spectacularly laying the foundations for modern psychometrics, while factor analysis was a dismal burden, consuming his energy and distracting his attention. Scholars may argue whether factor analysis wasted his time, but all agree he never returned to absolute scaling.

Thurstone's contributions to social science, however, go deeper than inventing modern psychometrics. His goal was an entirely new theoretical psychology based on instincts, needs, and aspirations "where the dynamic self finds overt expression" (1923, 356), "We should analyze . . . [human actions] . . . as the expression of cravings that originate in the organism and find particular modes of satisfaction in the stimuli that happen to be available" (1923, 368). In Thurstone's brave new cosmology, psychology studies the objective representation of these mental forces, his alternative to stimulus-response behaviorism and subconscious psychoanalysis.

His scaling methods conceptualized these mental forces as abstract linear continua, objectively measured on numerical scales, and their interrelations expressed as mathematical formulations. Thurstone's sweeping advance, the greatest single achievement rationalizing social experience since the Enlightenment, opened the door to a new science of mind, then stalled when he inexplicably succumbed to factor analysis. The ensuing dark cloud obscured both his measurement and psychology, costing him the momentum to advance psychology to an objective science. In 1954, at the end of his career, he expressed surprise over all the attention received by his difficult factor analytic techniques, while his simple measurement methods never became widely popular (1959, 15).

His most important works, those which promised a sound, objective basis for social research occurred within a short time. By the early 1920s, he had thought out the important conceptual issues for a new science which he discussed philosophically in three articles (1919a, 1923, & 1924). Then in 1925 he explained his new scaling method, quickly followed in 1927, 1928, and 1929 with clarification and elaboration. By 1930, it was over. His shift to primary mental abilities entangled him for years in methods incompatible with absolute scaling. The dark cloud drifted over 20th century social science as factor analysis fatally aroused the naïve enthusiasm of social researchers everywhere.

Volumes could be written about Thurstonian psychometrics: its central features, empirical benefits, and implications for advancing social research. Of course this story would start with his decisive rejection of classi-



Nikolaus Bezruczko

Nikolaus Bezruczko is originally from Linz, Austria and immigrated to the U.S. as an infant. He graduated from MESA, University of Chicago in 1990 and has published widely in child development, aesthetics, psychometrics, and education. He currently lives with wife Ambra Borgognoni Vimercati and stepdaughter Alice in Chicago and Rome, Italy. Favorite recreation is summer hiking in South Tyrolean Alps.

cal psychophysics, as well as raw scores and mental ages. Inconsistencies between Weberian and Fechnerian methods, limen determinations, and JND estimation instability are examples of psychophysical concepts Thurstone considered worthless to social research. To make this methodology meaningful, he needed to reconceptualize psychophysics. Instead of collecting perceptions of lifted weights and constructing a scale with physical units, he would identify distances between mental stimuli based on observer agreement with opinion statements using Fechnerian magnitude estimation methods. Then all he needed was a procedure for transforming ordered proportions into scale values and computing their error distributions. He would project mental structures on linear continua and model their quantitative properties with normal probability functions. Other improvements were also necessary, such as shifting from the method of equal appearing intervals to paired comparison, but the decisive step was to conceptualize a response continuum in terms of social objects such as attitude, opinion, or preference judgments. His ideas, however, were strange to psychologists and social researchers, and Thurstone faced enormous resistance and hostility. He tried to convince skeptics that subjective units were not only sensible and necessary, but easily estimated by selecting an arbitrary item on a continuum and using its error distribution as the scale unit. "The standard deviation of this dispersion for a standard stimulus could be chosen as a subjective unit of measurement." (1952, 307) His responses to objections included elaborate descriptions of his measurement philosophy in publications which fortunately now provide a detailed record of Thurstone's rationale for psychological measurement. Some main ideas are:

* ***Mental integrity.*** A mental integrity independent of overt behavior underlies the human tendency to engage in particular actions. Thurstone's defiant reaction to empty-headed Stimulus-Response psychology, this concept rationalizes an inferential approach to mental functioning.

* ***Discriminal process.*** An automatic perceptual process sorts the ambient flow of external stimuli to identify those that may be useful to the organism. Thurstone asserted they would show an error distribution on the stimulus continuum reproducing the subjective qualitative experience.

* ***Motive forces.*** A structure of motive forces lies dormant in the mental system. Its provocation by items reveals mental affinity toward particular stimuli and defines a psychological continuum. "They acquire conceptual linearity and measurability in the probability with which each of them may be expected to associate with any prescribed stimulus" (1927b, 51). "To the extent their probabilities of association with stimuli are nearly the same, to that extent will they tend to be adjacently spaced on the imaginary psychological continuum." (1927c, 419)

* ***Arbitrary units.*** Measuring in general is based on an arbitrary unit of measure whose practical usefulness is its linearity. Thurstone applied Fechner's JND technique to estimate unit measure on the subjective continuum.

* ***Absolute scaling.*** Social researchers grate at Thurstone's insight that scaling must be independent of the sample measured and unit of measure. (Many of them are still using raw scores/ratings, percentages, and grade equivalents.) "We have called the method absolute, not in the sense of measurement from an absolute origin but in the sense that the scale is independent of the unit selected for the raw scores and of the shape of the distribution of the raw scores" (1927c, 517). If the associational likelihood between any two points on the continuum "should be affected by the opinion of any individual person or group, then it would be impossible to compare the opinion distributions of two groups on the same base" (1928a, 417).

* ***Parameter linearity.*** "The sum of the subjective separations between the stimulus pairs AB/BC must be equal to the experimentally independent determination of the separation AC. If the continuum is unidimensional, then this simple type of check would establish the fact" (1952, 308). Referring to the additivity axiom in physical measurement, Thurstone presages probabilistic conjoint measurement for nonphysical observations.

* ***Item fit.*** Thurstone was explicit, scale items need both rational and empirical support. "The scaling method should be so designed that it will automatically throw out of the scale any opinion statements which do not belong in its natural sequence" (1928a, 417). Thurstone, however, did not support attempts to establish internal consistency coefficients for this purpose. In general, "correlation procedures constitute an acknowledgment of failure to rationalize the problem and to establish the functions that underlie the data" (1929a , 224). Thurstone was adamant, "correlation coefficients are symbols of defeat" (1929a, 240).

He developed a detailed methodology to apply these ideas. For example, Figure 1 taken from a 1926 article presents possibly the first cumulative item response curve ever published in a social research journal, now a standard presentation method. Figure 2 shows parallel item trace lines defining linear structure, the essential empirical evidence for a numerical variable. Another Thurstone contribution to social theory building is the variable map which positions item by person dynamics in a quantitative graphic structure. He considered the map an essential foundation for psychological theory and provided many examples. Figure 3 is his S and R continua, Thurstone's theoretical justification for generalizing psychophysics to nonphysical stimuli (1927a). In the 1920s, any objective, quantitative representation of social phenomena was an extraordinary achievement. Contemporaries such as Binet, Burt, and Thorndike were pioneering ability and

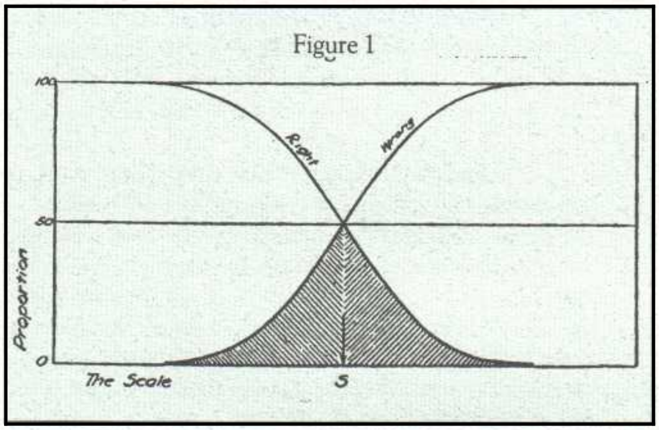


Figure 1

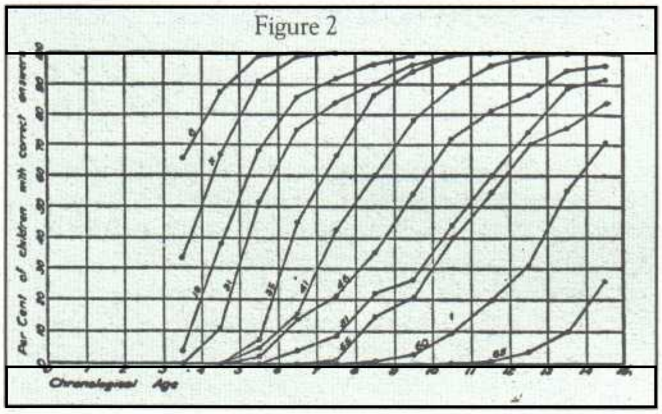


Figure 2

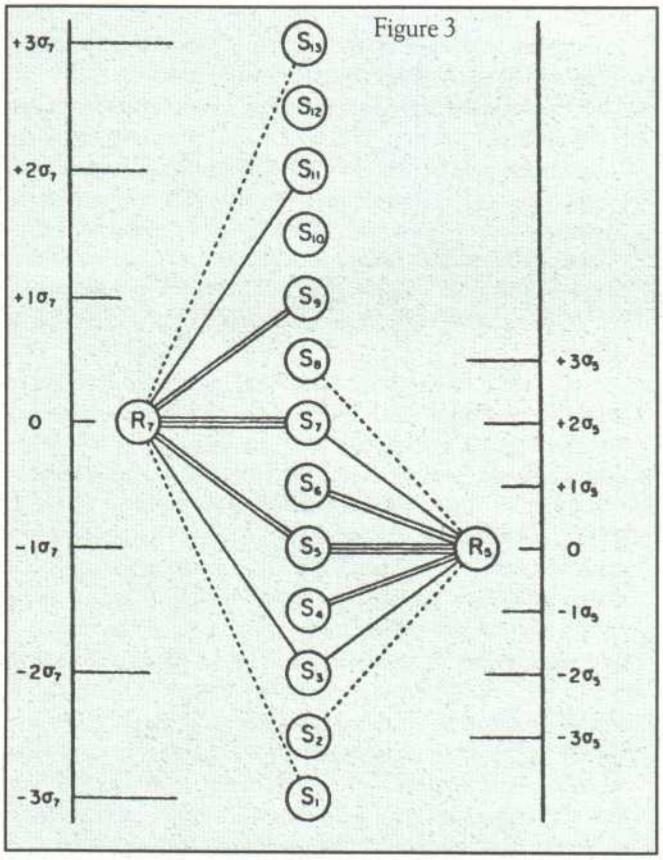


Figure 3

achievement tests, but no one commanded Thurstone's breathtaking view on a new science. Over the next 70 years, his ideas and methods would take on a life of their own ultimately to verify Thurstone's heretical assertion, "Attitudes Can Be Measured" (1928b).

In contemporary social research where hyper-quantification and over-parameterization are endemic, Thurstone is easily dismissed as a historical relic. After all, his whole scaling methodology is based on only two parameters, mean and standard deviation, the scale value and its error distribution. As we all know, the mathematical complications of contemporary social research far surpass Thurstonian methods. The surprise, however, is none of these complicated methods meet scientific rigor. Each of these highly touted methods (multidimensional scaling, cluster analysis, and so on), on close examination, suffers from critical defects that destroy its objectivity, generality, and simplicity. All of them obscure the person in data aggregation. While results are sometimes interesting, they are essentially descriptive techniques about specific samples. None offer any scientific advantages over Thurstonian measurement.

Newton's expression, "If I have seen farther than others, it has been by standing on the shoulders of giants" is appropriate here. The evolution of scientific methodology through Fechner to Thurstone and their successors, carries on an intellectual tradition over 4,500 years old as seen by the balance scales in Egyptian paintings during the Old Kingdom (Rice, 1990). We can only speculate how earlier cultures handled measuring issues. We know humans have an innate tendency to compare objects and abstract their differences. When commensurable with numbers and implemented to describe patterns of uniformity in nature, these units enable the scientific thinking responsible for Western civilization. Separating perceptual units from the observer and re-expressing their quantitative properties numerically is the milestone in human history underlying all abstract sciences. Commerce and its evolution into economics, for example, established social science. The failure of contemporary social research to continue this scientific methodology is responsible for its dismal record in the 20th century. Instead of modeling universal patterns, social research remains limited to fragmented, and inconsistent patterns of testimony, hardly scientific, generally failing to meet even minimal standards of replication or generality. (Some evidence suggests social research has degenerated to cult status, that is, dominated by obtuse methods which are only accessible to high priests yet without any clear relation to constructing scientific knowledge about human behavior.) The current absorption of social research by the physical and biological sciences is a commentary on this failure.

Thurstone provided the architecture for a new science of mind, as well as the foundations for a nonphysical measuring system: an objective framework in which to conduct scientific thinking. His key ideas are continuity, order, and variability. Continuity is the continuum underlying observations, order is the comparisons among items, and variability is the metric of precision. Georg Rasch, in turn, advanced objectivity by separating the ability and difficulty parameters. This achievement liberates social units from the confinement to standard deviates of arbitrary population means, and constructs a pure mathematical abstraction, a measured difference between ability and difficulty on an infinite continuum. Ben Wright advanced the framework even further by developing tests of statistical fit to detect departures of experience from the abstraction and improve precision and validity. Because this information about persons and items clarifies the dimensionality underlying a scale, it succeeds in eliminating the original motivation to develop factor analysis. Together they establish a measurement trilogy for the 20th century.

Biographical information concerning Louis Thurstone is documented in several sources (Guilford, 1957; Wood, 1962; Thurstone, 1952; see also Gulliksen, 1968). Thurstone was born in Chicago in 1887 to native Swedes, the Thunström family, who changed their name to Thurstone to accommodate American prejudice against foreigners. As a child, he was interested in music reinforced by his musician mother. As a teenager, he became interested in trigonometry and in college published an equation for trisecting any angle (1912). In 1912, he graduated from Cornell University with a mechanical engineering degree and immediately went to work for Thomas Edison in Orange, New Jersey (recruited after demonstrating his model of a nonflickering movie projector). In 1914, he started graduate school in psychology at the University of Chicago. While completing a learning function thesis, he went to Carnegie Institute of Technology in the Department of Applied Psychology. Thurstone returned to the Chicago Department of Psychology in 1924 where he founded the Psychometric Society and the journal *Psychometrika*. (Thurstone spoke on factor analysis to the Sigma Xi Society in spring 1948. After his talk, Ben Wright, then studying physics went to see Thurstone and learned from him his shortcut method for doing factor analysis by hand.) In 1952, Thurstone retired from the University of Chicago and moved his psychometric laboratory to the University of North Carolina. (References available on request.)

REFERENCES

Guilford, J. J. (1957) Louis Leon Thurstone: 1887- 1955. Biographical Memoirs, Volume 30. New York: Columbia University Press for the National Academy of Sciences.

Gulliksen, H. (1968). Louis Leon Thurstone: Experimental and mathematical psychologist. American Psychologist, 23, 716-80.

Rice, M. (1990). Egypt's Making: The Origins of Ancient Egypt 5000-2000 BC. London: Routledge.

Wood, D. A. (1962). Creative Thinker, Dedicated Teacher, Eminent Psychologist. Princeton, New Jersey: Educational Testing Service.

Thurstone, L. L. (1912). A curve which trisects any angle. Scientific American, 73, 259-261.

Thurstone, L. L. (1919a). The anticipatory aspect of consciousness. Journal of Philosophy, Psychology, and Scientific Methods, 16, 561-568.

Thurstone, L. L. (1919b). A scoring method for mental tests. Psychological Bulletin, 16, 235-240.

Thurstone, L. L. (1923). The stimulus-response fallacy in psychology. Psychological Review, 30, 354-369.

Thurstone, L. L. (1924). Contributions of Freudism to psychology: Influence of Freudism on theoretical psychology. Psychological Review, 31, 175-183.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-451.

Thurstone, L. L. (1926). The scoring of individual performance. Journal of Educational Psychology, 17, 446-457.

Thurstone, L. L. (1927a). Psychophysical analysis. American Journal of Psychology, 38, 368-389.

Thurstone, L. L. (1927b). A mental unit of measurement. Psychological Review, 34, 45-423.

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

Thurstone, L. L. (1928a). The measurement of opinion. Journal of Abnormal and Social Psychology, 22, 415-430.

Thurstone, L. L. (1928b). Attitudes can be measured. American Journal of Sociology, 33, 529-454

Thurstone, L. L. (1929a). Theory of attitude measurement. Psychological Review, 36, 222-241.

Thurstone, L. L. (1929b). The mental growth curve for the Binet Tests. Journal of Educational Psychology, 20, 569-583.

Thurstone, L. L. (1952). (Boring, E. G., Langfeld, H.S., Werner, H., & Yerkes, R,M., Eds.) A History of Psychology in Autobiography. 295-321., Worcester, Massachusetts: Clark University Press.

Thurstone, L. L. (1954). The measurement of values. Psychological Review, 61, 47-58.

Thurstone, L. L. (1959). The Measurement of Values. Chicago: University of Chicago Press.

Thurstone, L. L. & Chave, E. J. (1929). Theory of Attitude Measurement. Chicago: University of Chicago Press.

(See Mark Stone's article "Thurstone's Crime Scale Re-Visited" on page 53. Editor)

Jack Stenner:
The Lexile King

Linda J. Webster, Ph.D.

A. Jackson Stenner's accomplishments span academia and industry. Stenner is co-founder and CEO of MetaMetrics, Inc., a private corporation dedicated to educational research. His interest in measuring educational achievement began in a classroom in St. Louis at The Center of Our Lady of Grace.

"I taught emotionally disturbed children for three years. The Center took children who were too aggressive for public schools and who had various emotional problems. The kids were part of an in-patient six-to-ten week program where I taught during the day. I went to school at night. While there, I finished my undergraduate work and began the Master's degree."

With an undergraduate degree in psychology and education from the University of Missouri at St. Louis and a Master's underway, Stenner was awarded a Ford Foundation Fellowship that ran from 1970 through 1972. He moved to Washington, D.C. where he worked with the Council of the Great City Schools.

"This was my first look at measurement and policy. The Council represents the largest school districts in the U.S. by doing lobbying, policy analysis, and large-scale evaluation."

In 1972, Stenner left the fellowship program to start an agency that focused on social action research. "During the 1970's, we grew to over 250 people and worked with Head Start, the Department of Agriculture, and the National Career Evaluation Program. Head Start is a good example of what we were doing."

"When Congress mandated that the Head Start program be evaluated for its 'true effects', we spent several million dollars designing a study which would have been the most definitive study ever done. But, in the end, Congress chose not to fund the study."

Between 1973 and 1981 Stenner served as President and Director of NTS Research Corporation in Durham, North Carolina where, until the corporation was sold in 1981, Stenner was the Principal Investigator on educational research projects for the Food and Nutrition Service of the U.S. Department of Agriculture, the Administration for Children, Youth and Families for the Office of Human Development, the Washington, D.C. Public Schools, and the Office of Career Education.

During this period of NTS growth, Stenner was working on his Ph.D. at Duke University in Durham, North Carolina. By 1976, he had completed the course work, but, because of his busy research schedule, hadn't time enough to complete his dissertation until 1984.

Stenner's major achievements during these years was recognition of the fundamental importance in educational research of explicit construct specification, the empirical discovery that observable readability could be entirely predicted from word familiarity and sentence length and the application of this "Lexile Framework®" simultaneously to books and readers. "I worked on the Lexile Framework® primarily through grants from the NIH which funded twelve years of research on developing a better measurement system for reading and writing."

From 1984 through 1996, Stenner served as Principal Investigator on five grants from the National Institute of Health, all of which dealt with the measurement of literacy. And until 1996, he was also Chair and co-founder of the National Technology Group (NTG), a 700 person firm specializing in computer networking and systems integration.

Then in 1997, Stenner formed MetaMetrics, a company designed to make the Lexile Framework® easily available to schools and teachers, children and parents everywhere. "Our goal is to make the framework a global standard for language measurement. States are adopting the framework for their schools, tests and libraries: Hawaii, Utah, California, Alabama, North Carolina, and some parts of New York and Florida are in full swing."

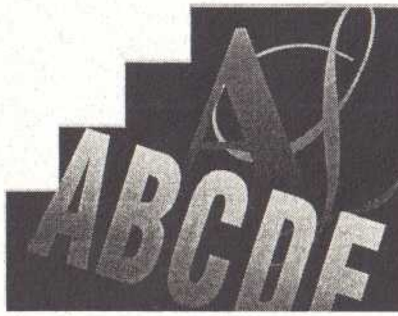
"The Lexile Framework® is an open standard that anyone can link to. Forty publishers use Lexiles as their means for building targeted products designed to bring readers and their books together in the most beneficial way. And more are coming on board every day."

In order to spread the availability and use of the framework, Stenner is on the road constantly, meeting and teaching with school boards, teacher organizations, politicians and business men as he works to help every school district in the nation to take advantage of the Lexile system for measuring books and readers on a common scale.

Of course there are critics. Some say it is all too complicated. Others insist it cannot be this simple. Each old guard must defend their turf. But the increasing number of publishers and school districts successfully basing the targeting of their products and teaching on the Lexiles system is gradually disarming most critics.

Stenner's research has appeared in many scholarly journals including *Popular Measurement, Rasch Measurement Transactions, Journal of Educational Measurement, Phi Delta Kappan*. Among the scholars collaborating with MetaMetrics are Dr. Donald Burdick of Duke University, Dr. Benjamin Wright of the University of Chicago and Dr. Ellu Page.

A. Jackson Stenner currently holds administrative or board positions with several professional organizations including: president of The Institute for Objective Measurement; board member for the North Caroline Electronics and Information Technologies Association (NCEITA), The National Institute for Statistical Sciences (NISS), and Duke Children's Hospital. He is also a member of the American Educational Research Association, the National Council on Measurement in Education, and Phi Kappa Delta.



The Lexile Framework® FOR READING

Educational Level	Literature Titles	Benchmarks	*Tests*/Textbooks
GRADUATE SCHOOL	**1700L** 1690 Concerning Civil Government 1680 Critique of Judgment 1660 On Abraham Lincoln 1660 On the Law Which Has Regulated the Introduction of New Species	**DISCOURSE ON THE METHOD AND MEDITATIONS ON FIRST PHILOSOPHY** To such a class of things pertains corporeal nature in general, and iis extension, the figure of extended things, their quantity or magnitude and number, as also the place in which they are, the time which measures their duration, and so on. That is possibly why our reasoning is not unjust when we conclude from this that Physics, Astronomy, Medicine and all other sciences which have as their end the consideration of composite things, are very dubious and uncertain; but that Arithmetic, Geometry and other sciences of that kind which only treat of things that are very simple and very general, without taking great trouble to ascertain whether they are actually existent or not, contain some measure of certainty and an element of the indubitable. *(Rene Descartes, author)*	1670 The Principles of Scientific Management; Dover Publications 1630 The American Constitution: Cases, comments, questions, 7th ed.; West Publishing 1610 The Condition of Postmodernity; Blackwell Publishers
	1600L 1570 Aeropagitica 1550 God, Idea of the Ancients 1540 History of Aeronautics 1530 Plutarch's Lives 1520 A Modest Proposal 1500 The Decameron	**FUNDAMENTAL PRINCIPLES OF THE METAPHYSICS OF MORALS** In fact, it is absolutely impossible to make out by experience with complete certainty a single case in which the maxim of an action, however right in itself, rested simply on moral grounds and on the conception of duty. Sometimes it happens that with the sharpest self-examination we can find nothing beside the moral principle of duty which could have been powerful enough to move us to this or that action and to so great a sacrifice; yet we cannot from this infer with certainty that it was not really some secret impulse of self-love, under the false appearance of duty, that was the actual determining cause of the will. *(Immanuel Kant, author)*	1550 Culture/Power/History: A Reader in Contemporary Social Theory; Princeton University Press 1530 On Injuries of the Head; Project Gutenberg 1510 On Human Nature; Howard University Press 1500 On Liberty; Hackett Publishing 1500 The Making of Memory: From Molecules to Mind; Doubleday
COLLEGE JUNIOR-SENIOR	**1500L** 1480 Eothen 1470 Utilitarianism 1450 The Prince 1440 The Legend of Sleepy Hollow 1420 Master Humphrey's Clock 1410 Aristotle's Physics	**ON ANCIENT MEDICINE** And as to him who had been accustomed to dinner, since, as soon as the body required food, and when the former meal was consumed, and he wanted refreshment, no new supply was furnished to it, he wastes and is consumed from want of food. For all the symptoms which I describe as befalling to this man I refer to want of food. And I also say that all men who, when in a state of health, remain for two or three days without food, experience the same unpleasant symptoms as those which I described in the case of him who had omitted to take dinner. *(Hippocrates, author)*	1450 Philosophical Essays; Hackett Publishing 1440 *Graduate Management Admission Test* GMAT 1430 *Certified Public Accountant Examination* CPA 1430 Criminal Justice Today; Prentice Hall 1410 Science and Education; The Citadel Press 1400 *Test of English as a Foreign Language* TOEFL
COLLEGE FRESHMAN-SOPHOMORE	**1400L** 1390 Moll Flanders 1350 Walden, or, Life in the Woods 1330 The Iliad 1330 Silas Marner 1320 Robinson Crusoe 1310 Up from Slavery	**THE SCARLET LETTER** But the point which drew all eyes, and, as it were, transfigured the wearer—so that both men and women who had been familiarly acquainted with Hester Prynne were now impressed as if they beheld her for the first time—was that SCARLET LETTER, so fantastically embroidered and illuminated upon her bosom. It had the effect of a spell, taking her out of the ordinary relations with humanity, and enclosing her in a sphere by herself. "She hath good skill at her needle, that's certain," remarked one of her female spectators; "but did ever a woman, before this brazen hussy, contrive such a way of showing it? Why, gossips, what is it but to laugh in the faces of our godly magistrates, and make a pride out of what they, worthy gentlemen, meant for a punishment?" *(Nathaniel Hawthorne, author)*	1390 *Graduate Record Examination* GRE 1380 *College Board Achievement Test in English* CBAT 1380 *Law School Admission Test* LSAT 1330 *Scholastic Aptitude Test* SAT 1330 *Medical College Admission Test* MCAT 1320 Psychology: An Introduction; Prentice Hall
TWELFTH GRADE	**1300L** 1280 Adam Bede 1280 From the Snow Image 1270 The Adventures of Robin Hood 1200 The Trumpeter of Krakow 1200 Great Expectations 1200 Civil Disobedience	**BROWN v. BOARD OF EDUCATION: 1954** Under that doctrine, equality of treatment is accorded when the races are provided substantially equal facilities, even though these facilities be separate. In the Delaware case, the Supreme Court of Delaware adhered to that doctrine, but ordered that the plaintiffs be admitted to the white schools because of their superiority to the Negro schools. The plaintiffs contend that segregated public schools are not "equal" and cannot be made "equal," and that hence they are deprived of the equal protection of the laws. Because of the obvious importance of the question presented, the Court took jurisdiction. Argument was heard in the 1952 Term, and reargument was heard this Term on certain questions propounded by the Court. *(347 US 483, 98 L ed 873, 74 S Ct 686)*	1290 Understanding Sociology; Glencoe/McGraw-Hill 1290 Speech Science Primer; Williams & Wilkins 1240 Business; Prentice Hall 1230 *Armed Services Vocational Aptitude Battery* ASVAB 1220 *Scholastic Reading Inventory* SRI-Level K 1210 *American College Testing Program* ACT
ELEVENTH GRADE TENTH GRADE	**1200L** 1190 Rebecca of Sunnybrook Farm 1190 Undying Glory 1180 Sense and Sensibility 1170 The Age of Innocence 1130 A Tale of Two Cities 1120 Agnes Grey	**WAR AND PEACE** Pierre had been educated abroad, and this reception at Anna Pavlovna's was the first he had attended in Russia. He knew that all the intellectual lights of Petersburg were gathered there and, like a child in a toyshop, did not know which way to look, afraid of missing any clever conversation that was to be heard. Seeing the self-confident and refined expression on the faces of those present he was always expecting to hear something very profound. At last he came up to Morio. Here the conversation seemed interesting and he stood waiting for an opportunity to express his own views, as young people are fond of doing. *(Leo Tolstoy, author)*	1160 History of a Free Nation; Glencoe/McGraw-Hill 1150 *NAEP Text* NAEP-Grade 12 1150 *Scholastic Reading Inventory* SRI-Level J 1130 America: Pathways to Present; Prentice Hall 1110 *Scholastic Reading Inventory* SRI-Level I
NINTH GRADE EIGHTH GRADE GRADE	**1100L** 1090 Antigone 1070 The Mystery of Edwin Drood 1070 All Things Bright and Beautiful 1020 Anne of Avonlea 1010 My Antonia 990 The Wizard of Oz	**PRIDE AND PREJUDICE** Occupied in observing Mr. Bingley's attentions to her sister, Elizabeth was far from suspecting that she was herself becoming an object of some interest in the eyes of his friend. Mr. Darcy had at first scarcely allowed her to be pretty; he had looked at her without admiration at the ball; and when they next met, he looked at her only to criticise. But no sooner had he made it clear to himself and his friends that she had hardly a good feature in her face, than he began to find it was rendered uncommonly intelligent by the beautiful expression of her dark eyes. *(Jane Austen, author)*	1090 *Scholastic Reading Inventory* SRI-Level H 1060 *Test of General Educational Development* GED 1050 *Test of Adult Basic Education, General Form* TABE-D 1010 *Scholastic Reading Inventory* SRI-Level G

The Lexile Framework

FOR READING

The Lexile Map

Educational Level: GRADUATE SCHOOL · COLLEGE JUNIOR-SENIOR · COLLEGE FRESHMAN-SOPHOMORE · TWELFTH GRADE · ELEVENTH GRADE · TENTH GRADE · NINTH GRADE · EIGHTH GRADE · SEVENTH GRADE · SIXTH GRADE · FIFTH GRADE · TH GRADE

Literature Titles	Benchmarks	Tests/Textbooks	
1700L	**DISCOURSE ON THE METHOD AND MEDITATIONS ON FIRST PHILOSOPHY**		
1690 Concerning Civil Government 1680 Critique of Judgment 1660 On Abraham Lincoln 1660 On the Law Which Has Regulated the Introduction of New Species	To such a class of things pertains corporeal nature in general, and its extension, the figure of extended things, their quantity or magnitude and number, as also the place in which they are, the time which measures their duration, and so on. That is possibly why our reasoning is not unjust when we conclude from this that Physics, Astronomy, Medicine and all other sciences which have as their end the consideration of composite things, are very dubious and uncertain; but that Arithmetic, Geometry and other sciences of that kind which only treat of things that are very simple and very general, without taking great trouble to ascertain whether they are actually existent or not, contain some measure of certainty and an element of the indubitable. *(Rene Descartes, author)*	1670 The Principles of Scientific Management; Dover Publications 1630 The American Constitution: Cases, comments, questions, 7th ed.; West Publishing 1610 The Condition of Postmodernity; Blackwell Publishers	
1600L	**FUNDAMENTAL PRINCIPLES OF THE METAPHYSICS OF MORALS**		
1570 Aeropagitica 1550 God, Idea of the Ancients 1540 History of Aeronautics 1530 Plutarch's Lives 1520 A Modest Proposal 1500 The Decameron	In fact, it is absolutely impossible to make out by experience with complete certainty a single case in which the maxim of an action, however right in itself, rested simply on moral grounds and on the conception of duty. Sometimes it happens that with the sharpest self-examination we can find nothing beside the moral principle of duty which could have been powerful enough to move us to this or that action and to so great a sacrifice; yet we cannot from this infer with certainty that it was not really some secret impulse of self-love, under the false appearance of duty, that was the actual determining cause of the will. *(Immanuel Kant, author)*	1550 Culture/Power/History: A Reader in Contemporary Social Theory; Princeton University Press 1530 On Injuries of the Head; Project Gutenberg 1510 On Human Nature; Howard University Press 1500 On Liberty; Hackett Publishing 1500 The Making of Memory: From Molecules to Mind; Doubleday	
1500L	**ON ANCIENT MEDICINE**		
1480 Eothen 1470 Utilitarianism 1450 The Prince 1440 The Legend of Sleepy Hollow 1420 Master Humphrey's Clock 1410 Aristotle's Physics	And as to him who had been accustomed to dinner, since, as soon as the body required food, and when the former meal was consumed, and he wanted refreshment, no new supply was furnished to it, he wastes and is consumed from want of food. For all the symptoms which I describe as befalling to this man I refer to want of food. And I also say that all men who, when in a state of health, remain for two or three days without food, experience the same unpleasant symptoms as those which I described in the case of him who had omitted to take dinner. *(Hippocrates, author)*	1450 Philosophical Essays; Hackett Publishing 1440 *Graduate Management Admission Test* 1430 *Certified Public Accountant Examination* 1430 Criminal Justice Today; Prentice Hall 1410 Science and Education; The Citadel Press 1400 *Test of English as a Foreign Language*	 *GMAT* *CPA* *TOEFL*
1400L	**THE SCARLET LETTER**		
1390 Moll Flanders 1350 Walden, or, Life in the Woods 1330 The Iliad 1330 Silas Marner 1320 Robinson Crusoe 1310 Up from Slavery	But the point which drew all eyes, and, as it were, transfigured the wearer—so that both men and women who had been familiarly acquainted with Hester Prynne were now impressed as if they beheld her for the first time—was that SCARLET LETTER, so fantastically embroidered and illuminated upon her bosom. It had the effect of a spell, taking her out of the ordinary relations with humanity, and enclosing her in a sphere by herself. "She hath good skill at her needle, that's certain," remarked one of her female spectators; "but did ever a woman, before this brazen hussy, contrive such a way of showing it? Why, gossips, what is it but to laugh in the faces of our godly magistrates, and make a pride out of what they, worthy gentlemen, meant for a punishment?" *(Nathaniel Hawthorne, author)*	1390 *Graduate Record Examination* 1380 *College Board Achievement Test in English* 1380 *Law School Admission Test* 1330 *Scholastic Aptitude Test* 1330 *Medical College Admission Test* 1320 Psychology: An Introduction; Prentice Hall	*GRE* *CBAT* *LSAT* *SAT* *MCAT*
1300L	**BROWN v. BOARD OF EDUCATION: 1954**		
1280 Adam Bede 1280 From the Snow Image 1270 The Adventures of Robin Hood 1200 The Trumpeter of Krakow 1200 Great Expectations 1200 Civil Disobedience	Under that doctrine, equality of treatment is accorded when the races are provided substantially equal facilities, even though these facilities be separate. In the Delaware case, the Supreme Court of Delaware adhered to that doctrine, but ordered that the plaintiffs be admitted to the white schools because of their superiority to the Negro schools. The plaintiffs contend that segregated public schools are not "equal" and cannot be made "equal," and that hence they are deprived of the equal protection of the laws. Because of the obvious importance of the question presented, the Court took jurisdiction. Argument was heard in the 1952 Term, and reargument was heard this Term on certain questions propounded by the Court. *(347 US 483, 98 L ed 873, 74 S Ct 686)*	1290 Understanding Sociology; Glencoe/McGraw-Hill 1290 Speech Science Primer; Williams & Wilkins 1240 Business; Prentice Hall 1230 *Armed Services Vocational Aptitude Battery* 1220 *Scholastic Reading Inventory* 1210 *American College Testing Program*	 *ASVAB* *SRI-Level K* *ACT*
1200L	**WAR AND PEACE**		
1190 Rebecca of Sunnybrook Farm 1190 Undying Glory 1180 Sense and Sensibility 1170 The Age of Innocence 1130 A Tale of Two Cities 1120 Agnes Grey	Pierre had been educated abroad, and this reception at Anna Pavlovna's was the first he had attended in Russia. He knew that all the intellectual lights of Petersburg were gathered there and, like a child in a toyshop, did not know which way to look, afraid of missing any clever conversation that was to be heard. Seeing the self-confident and refined expression on the faces of those present he was always expecting to hear something very profound. At last he came up to Morio. Here the conversation seemed interesting and he stood waiting for an opportunity to express his own views, as young people are fond of doing. *(Leo Tolstoy, author)*	1160 History of a Free Nation; Glencoe/McGraw-Hill 1150 *NAEP Text* 1150 *Scholastic Reading Inventory* 1130 America: Pathways to Present; Prentice Hall 1110 *Scholastic Reading Inventory*	 *NAEP-Grade 12* *SRI-Level J* *SRI-Level I*
1100L	**PRIDE AND PREJUDICE**		
1090 Antigone 1070 The Mystery of Edwin Drood 1070 All Things Bright and Beautiful 1020 Anne of Avonlea 1010 My Antonia 1000 The Wizard of Oz	Occupied in observing Mr. Bingley's attentions to her sister, Elizabeth was far from suspecting that she was herself becoming an object of some interest in the eyes of his friend. Mr. Darcy had at first scarcely allowed her to be pretty; he had looked at her without admiration at the ball; and when they next met, he looked at her only to criticise. But no sooner had he made it clear to himself and his friends that she had hardly a good feature in her face, than he began to find it was rendered uncommonly intelligent by the beautiful expression of her dark eyes. *(Jane Austen, author)*	1090 *Scholastic Reading Inventory* 1060 *Test of General Educational Development* 1050 *Test of Adult Basic Education, General Form* 1010 *Scholastic Reading Inventory*	*SRI-Level H* *GED* *TABE-D* *SRI-Level G*
1000L	**BLACK BEAUTY**		
960 Moccasin Trail 950 Secret Garden 940 Rosa Parks: My Story 930 The Grey King 920 Bonanza Girl 910 The Phantom of the Opera	One day, when there was a good deal of kicking, my mother whinnied to me to come to her, and then she said: "I wish you to pay attention to what I am going to say to you. The colts who live here are very good colts, but they are cart-horse colts, and of course they have not learned manners. You have been well-bred and well-born; your father has a great name in these parts, and your grandfather won the cup two years at the Newmarket races; your grandmother had the sweetest temper of any horse I ever knew, and I think you have never seen me kick or bite. I hope you will grow up gentle and good, and never learn bad ways; do your work with a good will, lift your feet up well when you trot, and never bite or kick even in play." *(Anna Sewell, author)*	990 *Scholastic Reading Inventory* 990 *NAEP Text* 940 World Cultures: A Global Mosaic; Prentice Hall 930 *Stanford Achievement Test* 910 *Test of Adult Basic Education* 900 *Stanford Achievement Test*	*SRI-Level F* *NAEP-Grade 8* *SAT 9-Advanced 2* *TABE-M* *SAT 9-Advanced 1*
900L	**TOM SWIFT IN THE LAND OF WONDERS**		
870 James and the Giant Peach 860 Julie of the Wolves 850 Titanic: The Long Night 830 Call It Courage 830 Frindle 810 My Side of the Mountain	Just what Tom's thoughts were, Ned, of course, could not guess. But by the flush that showed under the tan of his chum's cheeks the young financial secretary felt pretty certain that Tom was a bit apprehensive of the outcome of Professor Beecher's call on Mary Nestor. "So he is going to see her about 'something important,' Ned?" "That's what some members of his party called it." "And they're waiting here for him to join them?" "Yes. And it means waiting a week for another steamer. It must be something pretty important, don't you think, to cause Beecher to risk that delay in starting after the idol of gold?" "Important? Yes, I suppose so," assented Tom. *(Victor Appleton, author)*	870 Word 97; Glencoe/McGraw-Hill 870 *Scholastic Reading Inventory* 850 *Stanford Achievement Test* 820 *NAEP Text* 810 *Stanford Achievement Test* 800 *Scholastic Reading Inventory*	 *SRI-Level E* *SAT 9-Intermediate 3* *NAEP-Grade 4* *SAT 9-Intermediate 2* *SRI-Level D*
800L	**THE ADVENTURES OF PINOCCHIO**		
780 And Now Miguel 760 Gone-Away Lake 750 Pacific Crossing 740 Song of the Swallows 720 On the Banks of Plum Creek 700 My Name Is Brian	"Great soul!" said Pinocchio, fondly embracing his friend. Five months passed and the boys continued playing and enjoying themselves from morn till night, without ever seeing a book, or a desk, or a school. But, my children, there came a morning when Pinocchio awoke and found a great surprise awaiting him, a surprise which made him feel very unhappy, as you shall see. Everyone, at one time or another, has found some surprise awaiting him. Of the kind which Pinocchio had on that eventful morning of his life, there are but few. What was it? I will tell you, my dear little readers. On awakening, Pinocchio put his hand up to his head and there he found—Guess! He found that, during the night, his ears had grown at least ten full inches! *(Carlo Collodi, author)*	780 World Explorer: The U.S. & Canada; Prentice Hall 770 World Explorer: Latin America; Prentice Hall 760 World Explorer: Europe & Russia; Prentice Hall 760 *Stanford Achievement Test* 730 *Test of Adult Basic Education* 700 *Scholastic Reading Inventory*	 *SAT 9-Intermediate 1* *TABE-E* *SRI-Level C*
700L	**BUNNICULA: A RABBIT TALE OF MYSTERY**		

SEVENTH GRADE · SIXTH GRADE · FIFTH GRADE · FOURTH GRADE · THIRD GRADE · SECOND GRADE · FIRST GRADE

1000L

BLACK BEAUTY

960 Moccasin Trail
950 Secret Garden
940 Rosa Parks: My Story
930 The Grey King
920 Bonanza Girl
910 The Phantom of the Opera

One day, when there was a good deal of kicking, my mother whinnied to me to come to her, and then she said: "I wish you to pay attention to what I am going to say to you. The colts who live here are very good colts, but they are cart-horse colts, and of course they have not learned manners. You have been well-bred and well-born; your father has a great name in these parts, and your grandfather won the cup two years at the Newmarket races; your grandmother had the sweetest temper of any horse I ever knew, and I think you have never seen me kick or bite. I hope you will grow up gentle and good, and never learn bad ways; do your work with a good will, lift your feet up well when you trot, and never bite or kick even in play." *(Anna Sewell, author)*

990 *Scholastic Reading Inventory* *SRI-Level F*
990 *NAEP Text* *NAEP-Grade 8*
940 World Cultures: A Global Mosaic; Prentice Hall
930 *Stanford Achievement Test* *SAT 9-Advanced 2*
910 *Test of Adult Basic Education* *TABE-M*
900 *Stanford Achievement Test* *SAT 9-Advanced 1*

900L

TOM SWIFT IN THE LAND OF WONDERS

870 James and the Giant Peach
860 Julie of the Wolves
850 Titanic: The Long Night
830 Call It Courage
830 Frindle
810 My Side of the Mountain

Just what Tom's thoughts were, Ned, of course, could not guess. But by the flush that showed under the tan of his chum's cheeks the young financial secretary felt pretty certain that Tom was a bit apprehensive of the outcome of Professor Beecher's call on Mary Nestor. "So he is going to see her about 'something important,' Ned?" "That's what some members of his party called it." "And they're waiting here for him to join them?" "Yes. And it means waiting a week for another steamer. It must be something pretty important, don't you think, to cause Beecher to risk that delay in starting after the idol of gold?" "Important? Yes, I suppose so," assented Tom. *(Victor Appleton, author)*

870 Word 97; Glencoe/McGraw-Hill
870 *Scholastic Reading Inventory* *SRI-Level E*
850 *Stanford Achievement Test* *SAT 9-Intermediate 3*
820 *NAEP Text* *NAEP-Grade 4*
810 *Stanford Achievement Test* *SAT 9-Intermediate 2*
800 *Scholastic Reading Inventory* *SRI-Level D*

800L

THE ADVENTURES OF PINOCCHIO

780 And Now Miguel
760 Gone-Away Lake
750 Pacific Crossing
740 Song of the Swallows
720 On the Banks of Plum Creek
700 My Name Is Brian

"Great soul!" said Pinocchio, fondly embracing his friend. Five months passed and the boys continued playing and enjoying themselves from morn till night, without ever seeing a book, or a desk, or a school. But, my children, there came a morning when Pinocchio awoke and found a great surprise awaiting him, a surprise which made him feel very unhappy, as you shall see. Everyone, at one time or another, has found some surprise awaiting him. Of the kind which Pinocchio had on that eventful morning of his life, there are but few. What was it? I will tell you, my dear little readers. On awakening, Pinocchio put his hand up to his head and there he found—Guess! He found that, during the night, his ears had grown at least ten full inches! *(Carlo Collodi, author)*

780 World Explorer: The U.S. & Canada; Prentice Hall
770 World Explorer: Latin America; Prentice Hall
760 World Explorer: Europe & Russia; Prentice Hall
760 *Stanford Achievement Test* *SAT 9-Intermediate 1*
730 *Test of Adult Basic Education* *TABE-E*
700 *Scholastic Reading Inventory* *SRI-Level C*

700L

BUNNICULA: A RABBIT TALE OF MYSTERY

670 The Girl Who Loved Wild Horses
670 Amigo
660 Encyclopedia Brown Sets the Pace
620 M.C. Higgins, the Great
620 The Pizza Mystery
610 Beat the Story-Drum, Pum-Pum

"Of course he bites vegetables. All rabbits bite vegetables." "He bites them, Harold, but he does not eat them. That tomato was all white. What does that mean?" "It means that he paints vegetables?" I ventured. "It means he bites vegetables to make a hole in them; and then he sucks out all the juices." "But what about all the lettuce and carrots that Toby has been feeding him in his cage?" "Ah ha. What indeed!" Chester said. "Look at this!" Whereupon, he stuck his paw under the chair cushion and brought out with a flourish an assortment of strange white objects. Some of them looked like unironed handkerchiefs, and the others well, the others didn't look like anything I'd ever seen before. *(Deborah and James Howe, authors) Copyright © 1979 by James Howe. Reprinted by permission of Simon & Schuster Children's Publishing Division. All rights reserved.*

680 One Nation Many People, Volume One; Globe Fearon
670 Science (Grade 4); Addison-Wesley
650 *Test of Adult Basic Education, Anchor Test* *TABE-E*
630 Just Listen; Houghton Mifflin
600 Community Quilt; Scholastic Inc.

600L

A BABY SISTER FOR FRANCES

570 Curious George Takes a Job
550 Cousins
540 The Adventures of Sparrowboy
520 The Stories Julian Tells
520 John Henry: An American Legend
510 The Day Jimmy's Boa Ate the Wash

"Did you forget that I like raisins?" "No, I did not forget," said Mother, "but you finished up the raisins yesterday and I have not been out shopping yet." "Well," said Frances, "things are not very good around here anymore. No clothes to wear. No raisins for the oatmeal. I think maybe I'll run away." "Finish your breakfast," said Mother. "It is almost time for the school bus." "What time will dinner be tonight?" said Frances. "Half past six," said Mother. "Then I will have plenty of time to run away after dinner," said Frances, and she kissed her mother good-bye and went to school. After dinner that evening Frances packed her little knapsack very carefully. She put in her tiny special blanket and her alligator doll. *(Russell Hoban, author) Copyright © 1964 by Russell Hoban. Reprinted by permission of HarperCollins Publishers, Inc. All rights reserved.*

580 *Stanford Achievement Test* *SAT 9-Primary 3*
550 Communities; Harcourt Brace Jovanovich
540 People and Places; Silver Burdett Ginn
510 Team Spirit; Scholastic Inc.
500 Meeting Many People; Harcourt Brace
500 *Stanford Achievement Test* *SAT 9- Primary 2*

500L

THE MAGIC SCHOOL BUS INSIDE THE EARTH

490 Harold and the Purple Crayon
440 All Tutus Should Be Pink
420 Michael Bird-Boy
420 Angel Child, Dragon Child
410 Sam the Minuteman
400 Arthur's New Puppy

But suddenly, the bus began to spin like a top. That sort of thing doesn't happen on most class trips. When the spinning finally stopped, some things had changed. We all had on new clothes. The bus had turned into a steam shovel. And there were shovels and picks for every kid in the class. "Start digging!" yelled Ms. Frizzle. And we began making a huge hole right in the middle of the field. Before long CLUNK! we hit rock. The Friz handed out jackhammers. We began to break through the hard rock. "Hey, these rocks have stripes," said a kid. Ms. Frizzle explained that each stripe was a different kind of rock. We chipped off pieces of the rocks for our class rock collection. "These rocks are called sedimentary rocks, class," said Ms. Frizzle. *(Joanna Cole, author) THE MAGIC SCHOOL BUS is a registered trademark of Scholastic Inc. Copyright © 1987 by Joanna Cole. Reprinted by permission of Scholastic Inc. All rights reserved.*

480 *Scholastic Reading Inventory* *SRI-Level B*
480 Once Upon a Hippo; Scott Foresman
470 Bears Don't Go to School; Houghton Mifflin
440 Imagine That!; Scholastic Inc.
400 *Scholastic Reading Inventory* *SRI-Level A*

400L

FROG AND TOAD ARE FRIENDS

370 The Drinking Gourd
370 A My Name Is Alice
370 Owl at Home
360 The Best Way to Play
330 Clifford, the Small Red Puppy
320 Miss Nelson Is Back

"That button is thin. My button was thick." Toad put the thin button in his pocket. He was very angry. He jumped up and down and screamed, "The whole world is covered with buttons, and not one of them is mine!" Toad ran home and slammed the door. There, on the floor, he saw his white, four-holed, big, round, thick button. "Oh," said Toad. "It was here all the time. What a lot of trouble I have made for Frog." Toad took all of the buttons out of his pocket. He took his sewing box down from the shelf. Toad sewed the buttons all over his jacket. The next day Toad gave his jacket to Frog. Frog thought it was beautiful. He put it on and jumped for joy. *(Arnold Lobel, author) Copyright © 1970 by Arnold Lobel. Reprinted by permission of HarperCollins Publishers, Inc. All rights reserved.*

390 Discover Science (Grade 2); Scott Foresman
390 Carousels; Houghton Mifflin
350 Neighborhoods; Harcourt Brace Jovanovich
350 My World; Harcourt Brace
340 *Stanford Achievement Test* *SAT 9- Primary 1*
330 Who Painted the Porcupine Purple?; Silver Burdett Ginn

300L

CLIFFORD'S MANNERS

290 Sarah's Unicorn
270 Baseball Ballerina
270 In the Forest
260 At the Crossroads
230 The Boy Who Cried Wolf
220 Play Ball, Amelia Bedelia

Clifford loves to go visiting. When he visits his sister in the country, he always calls ahead. Clifford always arrives on time. Don't be late. Knock before you walk in. He knocks on the door before he enters. He wipes his feet first. Wipe your feet. Clifford kisses his sister. He shakes hands with her friend. Shake hands. Wash up before you eat. Clifford's sister has dinner ready. Clifford washes his hands before he eats. Clifford chews his food with his mouth closed. He never talks with his mouth full. Don't talk with your mouth full. Help clean up. Clifford helps with the clean-up. Say good-bye. Then he says thank you and good-bye to his sister and to his friend. Everyone loves Clifford's manners. *(Norman Bridwell, author) Copyright © 1972 by Norman Bridwell. Reprinted by permission of Scholastic Inc. All rights reserved.*

280 Too Big; Houghton Mifflin
270 *Test of Adult Basic Education* *TABE-L*
270 Parades; Houghton Mifflin
250 My Family, Your Family; Silver Burdett Ginn

200L

DANNY AND THE DINOSAUR

About The Lexile Framework* The Lexile Framework is a tool which helps teachers, parents and students locate challenging textbooks, literature titles and everyday world texts (like newspapers, periodicals and printed instructions). The Framework also allows determination of reader ability so that texts and reader may be appropriately matched. Text difficulty and reader ability are measured in the same unit: a Lexile*. A reader's measure is that position on the Lexile scale where the reader can expect to have 75% comprehension. Reader measures can be obtained from any test that has been linked to The Lexile Framework (Stanford Achievement Test, 9th ed., Scholastic Reading Inventory and the Stanford Diagnostic Reading Test). When reader ability measures match text difficulty measures, the reader is "targeted." Targeted readers experience confidence, competence and control over text and will want to self-engage in reading. Other factors (purpose, interest, developmental appropriateness, prior knowledge, text quality and text support) may be as important as the Lexile text measure when choosing a book for a reader. Please note that listed titles are illustrative only. Final determination of the appropriateness of a title rests with the user. The Lexile Framework Map is a component of The Lexile Framework, developed in part by a series of grants (HD 19448-01, HD 19448-02, HD 23430, HD 25358-01 and HD 25358-02) from the National Institute of Child Health and Human Development, National Institute of Health and United States Public Health Service. For more information about The Lexile Framework, contact MetaMetrics, Inc. at 1-888-LEXILES or www.lexile.com.

* "Lexile" and "Lexile Framework" are trademarks of MetaMetrics, Inc.

LEXILE TM

Look to the Lexile logo for appropriate reading levels...

READING RULER

Lexile Perspectives

A. Jackson Stenner, PhD and Benjamin D. Wright, PhD

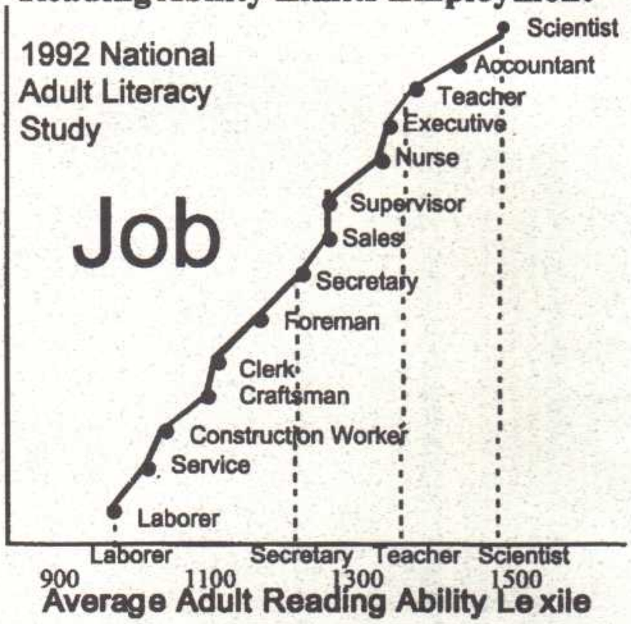
Job

In 1992, when 25,000 adults reported their jobs to the National Adult Literacy Study, their reading ability was also measured (Campbell et al, 1992; Kirsch, et al 1993, 1994). It turned out that the average laborer read at 1000 Lexiles, the average secretary at 1200, the average teacher at 1400 and the average scientist at 1500. Figure 1 summarizes this relationship between reading ability and employment.

There appears to be a correlation between an increased reading ability and improved job status that might prove to be of motivational value. Figure 1, makes an obvious statement that anyone wishing to be a teacher at 1400 Lexiles who reads at only 1000, must increase their ability by 400 Lexiles to reach that goal. In short, anyone serious about teaching might use the Lexile Framework® to determine where it is necessary to improve. A potential teacher who can take 1400 Lexile books off the shelf and read them easily knows that they can read well enough to be a teacher. But if that potential teacher finds him/herself at 1000 Lexiles, then they cannot avoid the fact that they are not yet ready to qualify for teaching; not until they master reading more difficult text.

Figure 1

Reading Ability Limits Employment

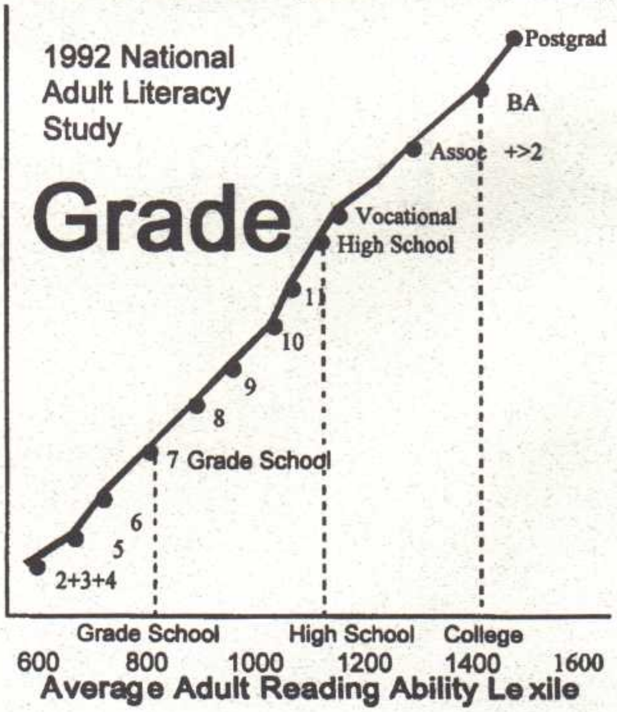


School

If we agree culturally that reading is learned in school, then the 1992 National Adult Reading Study shows that there is a strong relationship between the last school grade completed and subsequent adult reading ability. Figure 2 shows that, on average, we are never more literate than the day we left school. Therefore, the average 7th grade graduate reads at 800 Lexiles, the average high school graduate reads at 1150 Lexiles, and college graduates can reach 1400 Lexiles. The implication is that the last grade of school successfully completed defines one's reading ability for the rest of one's life; that once we leave school and we no longer benefit from the reading challenges that school provides, we tend to stop improving our reading abilities. The overwhelming implication of Figure 2 is that if we aspire to become a more literate society, then we must help everyone stay in school as long as it takes to achieve at some higher adult reading ability level.

Figure 2

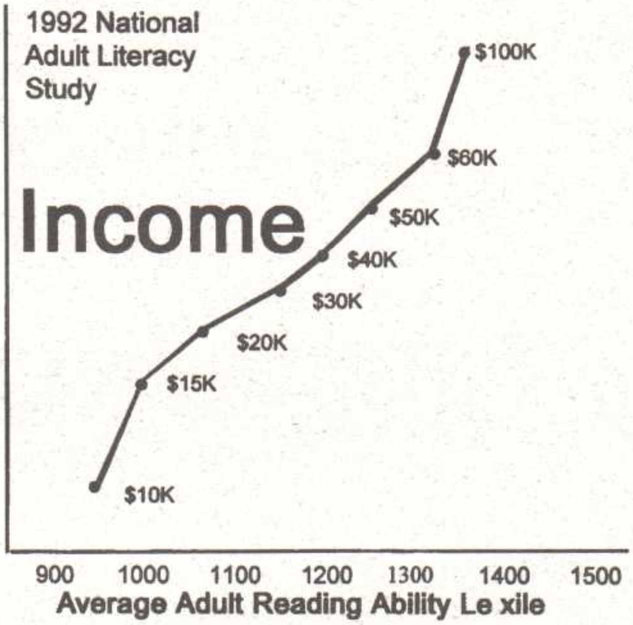
Leaving School Limits Reading Ability



Income

Using the data from the 1992 National Adult Literacy Study, it appears that reading ability is an indicator of how much we can expect to earn. Figure 3 shows the average incomes of readers at various Lexile reading abilities. From 1000 to 1300 Lexiles, each reading ability increase of 150 Lexiles doubles earning expectations. If one reads at 1000 Lexiles and wishes to double their potential, then they should attempt to improve their reading ability to 1150 Lexiles. When students can see the financial consequences of reading ability on an easy-to-understand scale that connects reading ability and income, then they have a persuasive reason to spend more time improving their reading abilities. The direct relationship of reading ability to income level illustrated in Figure 3 makes a strong argument that higher levels of reading ability should result in higher incomes, which might be used as a motivational tool when working with potential "drop-outs" or "stop-outs."

Figure 3
Reading Ability Limits Income



Reading Education

Education can succeed more fully if we connect learning to individual learner motives. If students feel engaged as individual learners, then perhaps it will be possible to engage their desires and arouse their drives. Engaged student education will drive itself, leaving us to add support and guidance. Otherwise, we will continue running a penitentiary system that keeps some troublesome kids off the street, but only for a while. When we know text readability, all we need to do to determine how well a student reads is to ask them to read a page or two aloud. If they succeed, we can give them a more difficult page. If not, we know their reading ability is below the readability of the text we asked them to read. No need for debate. No need for guesswork. No need for confusion or reproach. The student's status is plain to us and plain to them. We have not tricked them with a mysterious test score. All we have done is to help them see for themselves how able they are to read at specified levels of achievement.

Editor's Note: This is a reprint from last year. It is included again in this section to round out the Lexile Story.

> ***The Lexile Framework is a tool that has created a lot of excitement among our teachers. It's easy to use and has a great potential for impacting instruction.***
>
> Vickie Hugger
> C. B. Eller Principal
> Wilkes County, N.C.

How the Lexile Framework® Operates

Rick R. Smith

Matching readers and books

The Lexile Framework® enables teachers to develop personalized instruction based on Lexile measures and Lexiled reading lists. Properly targeted readers find reading a more entertaining and educational learning experience. Teachers and administrators worry about the need for individualized instruction. With test reporting Lexile measures and Lexiled texts, teachers, students and parents can, at last, design reading enrichment programs based on hard, scientific evidence. Readers who work in a Lexile reading program improve an average of 10 Lexiles per week.

Lexiles assist teachers and parents to plan educational strategies because the measures are not based on grade levels or age but on actual reading ability. Saying there is a fifth-grade reading level is like saying there's a fifth-grade shoe. We don't measure student's feet by age or grade. Why then would we be satisfied to measure their reading ability that way?

Readers and Books on a Same Scale

Methods for determining the readability of texts have existed for 50 years. Several are still in use. The Lexile Framework®, however, is different in three major areas:

(1) The analysis is based on the entire book. Every word is counted; every sentence length is recorded. Most other readability formulas are based on samples.

(2) Individual reading ability is measured by tests that measure on the same scale as the books are measured.

(3) The Lexile Framework® is an open standard that can be linked to any test, such as the Stanford 9 and its Diagnostic Reading Test.

The Lexile Framework® is like a thermometer. Our Celsius and Fahrenheit scales are absolute and easily equated through a simple formula. We rely on temperature measures to make decisions, to treat a cold, to determine what to wear. With the Lexile scale, we can determine the reading difficulty of texts and the reading ability of students. This puts students and books on the same scale enabling us to target a treatment method to each individual. By being able to put student measures on a scale of books, we can adjust their reading challenge. When a reader is bored, they don't work hard enough to benefit. When a reader is overwhelmed, they give up. But when their book is on target, then they read to the challenge and grow.

We would not expect first-year Spanish students to read Don Quixote de la Mancha. Too often we ask students to read texts they aren't ready for or books too far below their reading ability to be challenging. The Lexile Framework® solves that dilemma.

The Lexile Framework® determines the difficulty of virtually any text from its syntactic and semantic structure. Measures are assigned in ascending order of difficulty from the simplest children's books to *The New England Journal of Medicine*. The formula for calculating the level of difficulty is simple: difficulty is governed by two variables, the complexity of the syntactic structure and the vocabulary used.

Fify thousand books, including many world classics, have been measured. Among them are "Moby Dick" (Lexile measure 1210), "To Kill a Mockingbird" (920), "The Boxcar Children" (550) and Dr. Seuss' "One Fish, Two Fish. Red Fish, Blue Fish" (260). 150 additional books are added every week.

Lexile Tests

Text measure is only half of a comprehensive strategy to improve literacy. The second half is the placement of readers on the same scale through Lexile tests. A useful method for developing reading test items is "embedded completion". Passages from published works are used. Each test item has its own Lexile measure. When a student can understand the passage, their answer will be correct. These Lexiled items produce tests geared to the ability levels of students.

An easy Lexile item:

"The giant was mean. He was very ugly, too. We all ran away."

We were ____________
afraid
done
quite
tired

A more difficult example:

"Within several hours after launch, the Spacelab crew went to work on the experiments. To do work without stopping, the crew was divided into two shifts. Young, Parker and Merbold made up the red shift. Shaw, Garriott and Lichtenberg were the blue shift."

They ________ the work.
finished
graded
hated
shared

Each of the four options completes the sentence grammatically. However, only "afraid" and "shared" fit the passage. Following the administration of Lexile-linked tests, students receive a Lexile measure based on their pattern of correct responses. Typical Lexile tests contain 40 or more items.

Lexile Level

The next step is to direct students to books at their Lexile level. A decade of study by experts in measurement, testing and reading forecasts that students thrive when they can access approximately 75% of the materials they read. Students are therefore challenged, but not defeated, by a book set at their 75% success level. This is the basis for Lexile targeting.

In a typical fourth grade classroom, a teacher faces the challenge of developing a reading program for students at several different levels of ability. The Lexile Framework® program enables him or her to attack that problem as never before. For example:

Scholastic, Inc. has published biographies of Martin Luther King that are written at different levels of difficulty, different Lexiles. Students' Lexile measures can be used to assign the King book best targeted to their own reading ability. This is individualized instruction at its best. A teacher can assign 30 students to do a book report on Dr. King. Every student is studying the same subject but is doing it at his or her own level. This is the kind of classroom strategy that has a positive impact on the students and helps teachers who want to target students individually.

Lexile implementation has generated a lot of interest in our school system. Teachers are using the students' Lexile scores when developing classroom reading lists and/or providing supplemental books for unit studies.

Learning to effectively use Lexile scores has also supported the reading incentive program, "Accelerated Reader." Teachers are monitoring students' book selections so that the materials more closely match the student's Lexile level. As a system, our teachers are working on approprite sample book titles that fall within lexile ranges for Level I, II, III and IV students. This resource will be very helpful for instructional planning and also conferencing.

Judy Hall
K-5 Coordinator
Wilks County Schools

READING RULER

THE LEXILE COMMUNITY: *From Science to Practice*

Rick R. Smith

Lexiles in Education

Educators in Miami-Dade County, the nation's fourth largest school district were searching for tools to help them implement a strategy to improve literacy among their students. They needed a bridge between school and home in order to encourage parental involvement. Further, they wanted a private-public partnership which enlisted the help of the business community. The Lexile Framework® provided exactly what they needed and the Miami-Dade County school district is on its way to building the first Lexile-linked community.

The Lexile Framework® connects reading comprehension tests, books. magazines and newspapers to a common scale of measurement. With this tool, tests and reading materials are linked, and, as a result, the days of confusing norm-referenced test results that limit practical applications for students, parents and teachers are over. Teachers, parents, libraries and bookstores can use the Lexile scale to provide books that encourage and challenge readers at all levels in the best possible way.

The common framework joins schools, homes, libraries, publishers and bookstores in one shared Lexile Community. "This kind of public-private partnership can only be a boost to our efforts to improve literacy," said Ms. Norma B. Bossard, District Director of the Miami-Dade County School's Division of Language Arts and Reading.

A similar effort has been launched in Atlanta. "We want our children to love reading books, and we believe this comprehensive plan will help us achieve that goal," said Dr. Regina Johnson, the Language Arts Coordinator for the Atlanta schools. "We were impressed by the concept of Lexiles because of its method of linking testing and books. We need to diagnose our student's ability to read and then to assign relevant reading materials. Now we can use the books in our schools and libraries to meet their needs."

Standardized tests have not given teachers or parents the tools they need to change children's behavior. The results of standardized tests, be they percentiles or scores, are not useful. How can a child's percentile score help a teacher, parent or student to develop a plan for improvement when the percentile has no real-life application? With the Lexile Framework®, the student knows from their Lexile measure what materials they can read and where that measure puts them on the scale of educational development. Thus the Lexile Framework® gives schools a new and potent tool for encouraging and expediting parental involvement.

Three hundred thousand students were given spring Lexile-measured reading comprehension tests in Miami. The following summer, each student was encouraged to read 2 books from a recommended reading list geared to their respective reading levels. Students were tested again in the fall, and again in the spring to implement an aggressive student sensitive reading initiative.

Business and the Lexile Program

The schools aren't alone in this endeavor. Barnes & Noble and Books & Books provide the Lexiled books that are used in the Miami-Dade reading initiative. The world's largest book distributors, Baker & Taylor and Ingram Book Company, also participate, making sure that Lexiled books are in their inventory for rapid delivery. Miami has become a Lexile-linked community, where families, schools, book stores, libraries and business work together to make sure children have access to books that help them improve their reading.

Lexile communities are developing in North Carolina, California, St.Petersburg, Atlanta and many school districts in Kansas. Books and tests which utilize the Lexile Framework®, Scholastic Reading Counts! and Harcourt's reading comprehension tests (Stanford 9 and Metropolitan 8), are used in 18 of our largest school districts.

Lexiles are bringing together instruction and assessment - two worlds that in the past have been separate. This is crucial to building a community which implements programs that help children improve their reading. The assessment tools are Lexile-linked reading comprehension tests. The instructional tools are Lexile-based reading lists. Now that the book industry is involved, parents and students can get measured reading lists and targeted books that maximize their children's opportunity to improve their reading.

The Framework Spreads

Miami-Dade and Atlanta are only 2 examples of the growing acceptance of the Lexile Framework® across the educational marketplace. Today the number of students who receive Lexile reading measures and Lexiled reading lists exceeds 10 million. The use of the Lexile Framework® is spreading. Lexiles are on the way to becoming the standard by which the measurement of reading comprehension and the determination text and book readability are measured.

The United States Department of Education endorses the framework for its "America Reads" program. California measures its reading list in Lexiles. Harcourt-Brace Educational Measurement reports the results of its standardized tests in Lexile measures. Scholastic, Inc. has lexiled 4,000 titles. Texas, Kentucky, New York and Boston are adopting Lexile-measured products.

Lexiles In Miami

The Lexile Framework® is essential to the Miami-Dade County comprehensive reading plan. The plan requires that students, after taking Scholastic Reading Inventory tests which report Lexile measures, read five books during each nine-week grading period, plus additional books during the summer. Books must be Lexiled to be on the district's reading list. Each Miami-Dade County Public School student in grades 2-11 is given a Lexile reading test to determine their reading level and to plan their instruction.

Lexiles in Atlanta

Atlanta uses a Lexile program for 30,000 students in grades 1 through 5. Students receive Scholastic Reading Inventory Lexile measures that help them shape their own reading programs. Atlanta gets the right book to the right student at the right time.

Barnes & Noble, Baker & Taylor and Ingram Book Company are ensuring that the Lexiled books on recommended reading lists are readily available in Atlanta stores and libraries. Education is the number one concern of the American people. A major reason for this concern is poor literacy skills. This public-private partnership is replicating around the country. When schools, libraries and industry can go into battle together against a common foe, they will succeed.

Lexiles in California

California has Lexiled 500 books for its State of California reading list. Harcourt Brace has linked its Stanford 9 and Stanford Diagnostic Reading Test to the Lexile Framework®. This means that California's 5.9 million students receive Lexile measures that show them their positions in the Lexile Framework® and hence the books that are best for developing their reading. California school districts provide a targeted reading list to each of their students, based on each student's test results.

Lexiles in North Carolina

North Carolina students take Lexile-linked reading comprehension tests and receive reading lists called Pathfinders. These lists are drawn from a bank of 6,000 Lexiled titles. End-of-grade Lexile measures enable students, parents and teachers to design individually targeted summer reading programs which are not guesses or hunches but specific and positive educational interventions.

Lou Fabrizio of the North Carolina Department of Public Instruction, said that reporting test results in Lexiles makes the measures relevant because now parents, students and teachers can use them to determine a specific plan of action to improve reading. "The reason why we use Lexiles now is because one of the biggest complaints we get about test results is that no one knows what do you do with them when they get them back."

Local school districts in North Carolina are using the Lexile Framework® in a variety of ways. Craven and Gaston County use the framework to align their instructional programs, such as Accelerated Reader, with the North Carolina End-of-Grade test. Prior to the Lexile Framework®, teachers and media coordinators had to rely on the reading levels suggested by individual publishers. Since no common framework existed, schools were never sure whether the instructional materials ordered would turn out to be at the correct levels. With the Lexile Framework®, teachers and librarians can adjust the materials they already have to levels that align with the state assessment.

Lexiles in "America Reads Now!"

The US Department of Education's "America Reads Now" program recommends Lexile-rated books. Secretary of Education Richard Riley uses the framework in the department's "Checkpoints for Progress" manuals to guide tutors, parents and teachers in their work with students. The million tutors who work in the program are given guides which recommend how to improve reading and explain the Lexile Framework®. The books recommended for reading are also Lexiled.

Lexiles in Spanish

A Spanish Lexile Framework® has been developed and is on its way into educational practice. Frameworks in French and German are underway. The Spanish framework was developed at the request of states and companies that wanted to meet the needs of students and workers in homes where English is not the first language. The ultimate goal is to build a universal metric for the measurement of readability in any language, equated across languages. This will enable people learning a new language to measure their level of comprehension in a language specific framework of Lexiled tests and Lexiled texts which is further equated across languages. This will enable each student to design the individual instructional program which best helps them reach the level of fluency they require - for whatever reasons.

Best Practices for Using Lexiles

Barbara R. Blackburn

In today's educational climate, cries for quick fixes and immediate solutions are endless. Lists of "best practices" abound, and reform often means jumping on the latest bandwagon and expecting major changes immediately. This approach results in what has been described as "Teflon education"-guaranteed not to stick. As educators, the ineffectiveness of this approach calls for a different design. Therefore, when looking at "best practice" using the Lexile Framework®, it is critical to set specific parameters. The strength of the Lexile Framework® is its flexibility in terms of use, but the Framework can be misused because of a lack of understanding of its purpose.

The Lexile Framework® is a tool for looking at reader ability relative to the difficulty of text. It allows a parent, student, teacher, or media coordinator to understand the performance of a reader (whether on a standardized test or informal assessment) through examples of text materials (books, newspapers, or magazines) the reader can understand, rather than through a number such as a stanine or percentile. While the ability to link student performance on a test or other assessment tool with text materials is a powerful tool, the major misconception regarding the Lexile Framework® is that the framework is a program or method for teaching students to read. Rather, the Lexile Framework® is a tool that can be used with existing programs, methods, and strategies to enhance reading growth. Using the framework in the most effective manner means starting with the realization that it does not replace any program a school may be currently using nor is it a way to actually teach reading. It is a tool — a knowledge base — that can enhance reading methods and sharpen the focus of instructional programs currently in use in a school or district.

Barbara Blackburn
Blackburn Consulting Group

The Lexile Framework® provides:

1. A way to define (with books and other text materials) what is above grade level, on grade level, and below grade level, according to the standardized test used.
2. A way to understand a student's location on the reading spectrum, based on their performance on a standardized test or informal assessment.
3. A way to match classroom libraries, resource materials, textbooks, and library materials to standardized tests.

Several districts in North Carolina have been using the Lexile Framework® to enhance their current programs and to more sharply align their instruction with the state assessment, or End-of-Grade Tests (EOG). A foundational use of the framework begins with using it to understand the EOG. What must a student be able to do to score "on grade level" on the EOG? For math, the answer was simple. Clear-cut, concrete objectives were provided and teachers had stable benchmarks for achievement. Reading, on the other hand, was not simple. Clearly, students must be able to answer certain types and levels of questions and they should be able to read "on grade level", but what does that mean? If, as a fourth grade teacher, my students can read the state approved fourth grade textbook and answer questions, is that enough?

The Lexile Framework®, when introduced in North Carolina, answered that question. Students' scores on the EOG are converted into lexiles. In addition to providing diagnostic information for each student, teachers could now take the students who scored at level 3 (state designation for grade level), see the Lexile range for that level, and have an estimated idea of "grade level" text materials. Benchmarking books and other text materials at "grade level" provided a starting point for structure for the reading portion of the test.

The application of this information is immediate. Simply by knowing where specific book titles fall in relation to the EOG, teachers have a way to evaluate the appropriateness of those books used in the classroom. For example, many fourth grade teachers use the novel "Tales of a Fourth Grade Nothing", (490 lexiles). 490 lexiles is well below the level 3 (on grade level) range of 625-880 lexiles at fourth grade on the EOG. Although this text is an age-appropriate selection, it is not a book that appropriately challenges students on grade level in light of the EOG. While this does not mean that "Tales of a Fourth Grade Nothing" is an inappropriate book for fluent, easy, pleasure reading, it does indicate that the text should not be used for a significant portion of instructional time. For a teacher, with all the pressing needs and curriculum objectives to cover, it is critical to focus and align instructional materials appropriately, particularly in regard to state and national standards and accountability tools.

How are schools and school districts using the Lexile Framework® effectively? Evaluating current resources and aligning their use to match accountability measures is one of the strongest instructional uses of the Lexile Framework®. Craven County, North Carolina, is a case that illustrates the problems in assuming accountability measures. Each individual school had a variety of books and other text materials from a large range of publishers. Publishers provided recommended grade levels for each book, but there seemed to be a lack of consistency in the levels. Some books even had different levels, depending on the publisher or book list referred to. Several years ago a great deal of emphasis, time, and money, had been placed on a commercial, computerized program that most schools in the county implemented to provide a base leveling system and some consistency. However, reading test scores in the district were not improving at the rate desired. The vendor's marketing materials claimed the program appropriately targeted readers for growth, but this was not happening. Growth was shown on the commercially-provided test bank, but it wasn't transferring to the EOG. A portion of the problem was the leveling system used. Based on a combination of readability formulas, the computer system relied on grade equivalents. The underlying assumption was that since everyone defines a grade level the same way, a simple grade equivalency can be used. However, there was no way to know if the grade equivalents matched the state testing definitions of "on grade level". Enter the Lexile Framework®.

Using a comparison database of the grade levels and Lexile levels, over 6,000 books could be evaluated to see if the grade levels actually matched the state levels. Although many did, a large number of title levels did not match the test (see Table A on page 24). In fact, many books that were leveled at a particular grade level were actually considered level two (or below grade level) according to the EOG Lexile score data.

The result was that many students were reading books considered "on grade level", but these books were actually easier than the appropriate level of difficulty for the state assessment. This explained part of the lack of growth on the EOG. However, the district was not forced to choose between their computer program and Lexiles. Because the Lexile Framework® is a tool, they simply began to use the Lexile Framework® to adjust and customize the computer program to meet their needs. Teachers, parents, and media specialists could simply direct students to other choices, that are more challenging. The issue is not that a student shouldn't be allowed to read easy books. But for growth, there must be a balance of easy, fluent reading, and reading that is appropriately challenging. In this case, everyone assumed the books were challenging (based on the levels provided), when they weren't. As a librarian in Gaston County, NC noted, "No wonder our brightest students aren't growing. They are reading books we thought were harder, but in reality, they're not!" And as one principal said, "We don't need any help picking easy books. Students do that on their own." In several of the districts in North Carolina, teachers and media specialists are using the computerized comparison to better target appropriately challenging books that match the state ranges of performance.

Best practice, however, moves past simply aligning curricular resources with assessment. It also uses assessment to inform instruction. A special education teacher in Wilkes County, NC used the comparison of the popular software program in a different way. One of her students was desperate to read a book that was "on grade level" and had "the right number of points." Unfortunately, he was performing well below grade level, and was struggling to find a book he could read that was also popular with his peer group. The teacher used the computerized comparison to find titles that were "grade level" but were actually much easier (such as "Fourth Grade Rates" in Table A). She directed the student to selected books at his Lexile range that also were leveled (and labeled in this case) at a higher grade level. In effect, she turned a negative (books leveled incorrectly in light of the EOG) into a positive for her student.

Another way the Lexile Framework® can allow a teacher to customize instruction is to modify the traditional class novel. In a typical classroom, if all students read one novel, it is probably easy for some students, hard for a portion of the class, and right on target for the middle group. Depending on the ability range of students in the class, one novel probably is appropriate for 30-50% of the class. An alternative to this is using several novels, tied together by theme or genre. For example, in a fifth grade class, instead of everyone reading "Hatchet" by Gary Paulsen, students could be placed into literature circles of four to six students, based on Lexile levels. Then, each literature circle could choose a book by Gary Paulsen that falls within 50-100 Lexiles of their range. The teacher moves around the class to facilitate the small group discussions, but then pulls the entire class back together for an author study, which includes a comparison of different Gary Paulsen books. By using flexible grouping and a variety of titles, the teacher provides reading materials for each student at his or her ability level, but balances the instruction (and avoids tracking students) with the whole class activities. Similarly, if assigning book reports to a class, providing a range of titles within a genre, such as biographies, insures that students are provided opportunities to read material that is appropriately challenging to each individual. Unfortunately, far too often, students are left to pick books on their own, with no direction. Many students pick the easiest book they can find, and others are left hopelessly overwhelmed by books far above their level. Providing lists of books to students (and parents) that are linked to their Lexile level strengthens the chances of choosing books that are appropriate for growth.

The Lexile Framework® provides endless possibilities for use in schools, and this review only begins the discussion. Links between the media center and the classroom, between the public library and the school, between parents and teachers are easily forged using the framework. However, the most effective "best practice" instructionally with the Lexile Framework® is to evaluate one's current instructional practices, disaggregate available student data, and work with a curriculum consultant to determine the best way to use Lexiles in a particular situation.

Table A:
Sample Comparison of
Commercial Software Program and EOG
Grade Four: EOG Level 3 (on grade level)-625-880L

Title	Program Level	Lexile Level
"Fourth Grade Rates"	4.0	340
"Trumpet of the Swan"	4.1	860
"Jip: His Story"	4.2	860
"George Washington"	4.2	510
"Who Stole the Wizard of Oz?"	4.3	520
"Soup"	4.5	740
"Cherokee Indians"	4.6	390
"Tuck Triumphant"	4.8	850
Wayside School is Falling Down	4.9	440
"The Cybil War"	4.9	730

A Spanish Version of The Lexile Framework® for Reading

Ellie Sanford

"Accountability for student performance is one of the two or three-if not the most-prominent issues in policy at the state and local levels right now," stated Richard F. Elmore, a professor at Harvard University's Graduate School of Education (Olsen, 1999). Based on the survey conducted as part of *Quality Counts* '99, 48 states now test their students, and 36 publish annual report cards on individual schools.

Many states require students identified as limited English proficient to take the same tests as fully proficient students. While this may work for school or district accountability, it does not help these students to improve their reading skills. Some states allow these students to be exempt from the assessment for a limited period of time e.g., two years. But, the policies often require that the schools "adopt appropriate evaluative standards for measuring the progress of limited English proficient students in school" (North Carolina State Board of Education, Policy ID Number HAS-K-000).

The question often asked is "What reading skills does the student need to work on and what has been mastered?" That question deals with the reading skills the student actually possesses regardless of the language that the material is presented in. The skills needed to be a proficient reader in English-identifying, selecting, and collecting information; analyzing, synthesizing, and organizing information and discovering related ideas, concepts, or generalizations; and applying, extending, and expanding on information and concepts-are the same skills needed to be a proficient reader in any language. The only difference is the language that the material is presented in.

Readability equations can be used to order text in terms of comprehensibility. Likewise, reading tests can be used to order readers in reading skills. What distinguishes the Lexile Framework® is its ability to conjointly order texts and readers on the same scale. The ability to characterize a reader as 1000L and a text as 1000L enables a forecast of the comprehension rate that the reader will experience with that particular text. Comprehension, itself, is not an absolute; rather it is the consequence of an encounter between a reader and the text. The Lexile Framework® provides a single scale that can be used for targeting readers with text that provides an appropriate level of challenge. [For further information concerning The Lexile Framework® refer to the following documents: Scholastic Inc., 1999; Stenner, 1996; Stenner and Burdick, 1997; and Wright and Stenner, 1999.]

In 1998, MetaMetrics, Inc. undertook research to apply the premise that reading skills are independent of the presentation language. This project began with the development of a scale comparable to the Lexile Framework® that could be used to estimate the readability of Spanish texts and the reading ability of Spanish readers.

What did we do to develop a Spanish readability equation?

The first step in developing the Spanish readability equation was to identify English items that had confirmed the Lexile Theory. Differences between theoretical measure and empirical measure was small; less than 90L. The Lexile calibrations of the 227 selected items ranged from 260L to 1420L. Next, the 227 items were translated into Spanish for meaning, not just literal translations. Three items were not used because they did not work in Spanish e.g., a passage about the differences between "to," "too," and "two". The remaining 224 items were then translated back into English by a different set of translators.

The third step was to evaluate the accuracy of the translation process. The original English version of each item was compared with the back-translated version to identify those items that did not lose their meaning in the translation process. Five reviewers examined both versions of each item. An item was retained if the overall meaning remained the same and the statement could still be answered. In addition, the foils for each item were examined to see if they were still at the same level of difficulty. A total of 133 items were retained for further analyses.

The next step was to examine the text features that related to the difficulty of the Spanish items. All symbol systems share two features: a semantic component and a syntactic component. In language, the semantic units are words. Words are organized according to rules of syntax into thought units and sentences (Carver, 1974). Semantic units vary in familiarity and syntactic structures vary in complexity. The comprehensibility or difficulty of a message is dominated by the familiarity of the semantic units and by the complexity of the syntactic structures used in the message.

For the semantic component, it is clear that operationalization is a proxy for the probability that an individual will encounter a word in a familiar context and thus be able to infer its meaning (Bormuth, 1966). The semantic difficulty of Spanish text was estimated by calculating the mean of the log word frequency of each word in the text. The word frequency measure used was the raw count of how often a given word appeared in a corpus of 3,981,128 words sampled from a broad range of topics.

In the English Lexile Framework®, the syntactic complexity of a text is estimated by calculating the mean number of words per sentence in the text. Specific editing rules are employed to adjust for one-word sentences and dialogue qualifiers e.g., "said Patrick" and "Ami said." In English, dialogue qualifiers with two or less words are appended to the previous sentence (for example, " 'I see the moon,' he said." would be treated as one sentence, whereas, " 'I want to go to the store,' John stated loudly." would be treated as two sentences).

The same rules used to determine sentence length in English were used with Spanish texts except in the case of dialogue. In Spanish, dialogue qualifiers with three or less words were appended to the previous sentence.

Next, a regression analysis used the Spanish semantic and syntactic characteristics of the item to predict the reading comprehension difficulty of the 133 items in English. The premise was that overall comprehension difficulty of text is language independent. Four variables were used to quantify the difficulty of the text in English: (1) the theoretical Lexile measure of the original text, (2) the empirical Lexile measure of the original text, (3) the theoretical Lexile measure of the back-translated text, and (4) the mean theoretical Lexile measure of the text. The four analyses resulted in R^2s of greater than 0.89 and RMSEs less than 84L.

The mean difference between the original theoretical Lexile measures of the items and the back-translated Lexile measures of the items was 24.17L (N = 133 items). This process involved two sets of translations (English to Spanish and then back to English). In order to go from English to Spanish only one translation is needed. Therefore, the difference between the original English Lexile measures of the items and the mean theoretical Lexile measures of the items (original and back-translated) corresponds to the amount associated with one translation ($0.5 \times 24.17 = 12.085$). The final regression equation was derived from the Spanish semantic and syntactic characteristics (independent measures) of the 133 items and the mean theoretical Lexile measure of the English item (criterion measure). This equation explained most of the variance found in the set of reading comprehension items ($R^2 = 0.938$).

Validation of the Spanish Lexile Framework® is being examined from two perspectives: the text and the reader. The text perspective is being examined by looking at the level of difficulty of matched texts e.g., newspapers, literature, and empirical difficulty of items administered to native Spanish-language readers and basal readers. The reader perspective is being examined by looking at the relationship between level of reading comprehension and growth of native Spanish-language readers (Puerto Rico public and private school students), other standardized measures of reading comprehension, and teacher judgements of reading comprehension level.

How will the Spanish version of the Lexile Framework® be used?

MetaMetrics is developing the following materials for the classroom: (1) Spanish Lexile Framework® Map with representative titles and authors from across the Spanish-speaking world; (2) a series of assessments for students in grades 1 through high school to evaluate a student's reading comprehension skills when English is not their primary language; and (3) a series of Reading Pathfinder lists to be used with Spanish-speaking students to identify texts that match their reading comprehension level to instill more reading.

Not all languages are the same!

During this research we learned about differences between the structures of Spanish and English. It was hard to develop a corpus of Spanish text that could be used to construct the word frequency measure. Many Spanish books are actually translations of English books. It was much harder to find text that was originally written in Spanish.

The average length of Spanish sentences is longer than English sentences and the average length of Spanish words is longer than English words. This impacts readability formulas that use word length. Another difference between English and Spanish is word usage, e.g., verb tenses and masculine/feminine versions of the same word. Also, dialogue in Spanish differs from dialogue in English in the markers used, the placement of markers, and the length of qualifiers.

References

Bormuth, J.R. (1966). Readability: New approach. Reading Research Quarterly, 7, 79-132.

Carver, R.P. (1974). Measuring the primary effect of reading: Reading storage technique, understanding judgments and cloze. Journal of Reading Behavior, 6, 249-274.

Education Week. (1999). Quality Counts '99: Executive Summary-Demanding Results. Education Week on the WEB, 18(17), 5. URL: www.edweek.org/sreports/qc99/exsum.htm.

Olsen, L. (1999). Quality Counts '99: Shining a spotlight on results. Education Week on the WEB, 18(17), 8. URL: www.edweek.org/sreports/qc99/ac/mc/mc-intro.htm.

Scholastic Inc. (1999). Scholastic Reading Inventory, Technical Report #1. New York: Author.

Stenner, A.J. (1996, October). Measuring reading comprehension with the Lexile Framework®. Paper presented at the California Comparability Symposium, Burlingame, CA.

Stenner, A.J. & Burdick, D.S. (1997, January). The objective measurement of reading comprehension in response to technical questions raised by the California Department of Education Technical Study Group. Durham, NC: MetaMetrics, Inc.

Wright, B.D. & Stenner, A.J. (1999). Reading Ruler. Popular Measurement, 2(1), 34-42.

MEASUREMENT MUSINGS

The Shift from Modeling Observations to Applying Theory:

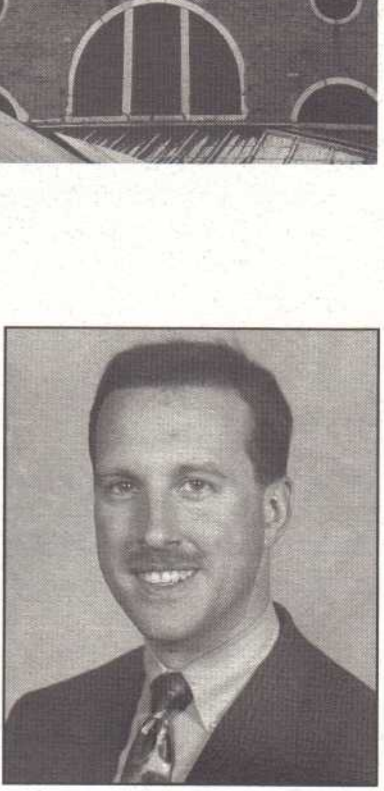
Some Timely Points About Measuring Latent Traits

Thomas R. O'Neill

American Dental Association

Tremendous progress has been made in the physical sciences in the last 500 years and the rate of change has been increasing, especially over the last 100 years. The discovery that some traits are transmitted genetically has led to the genetic engineering of fruits and vegetables, the cloning of mammals, and the promise of successful genetically engineered solutions to medical prob-lems. Military technology has been changed not only by the invention of the mass-produced rifle, but also by the radio, microchips, satellites, airplanes, and missiles which can deliver explosives or non-conventional weapons (chemical, biological, or nuclear). Medical technology has been changed not only by the invention of antibiotics, anesthesia and the development of a germ theory of disease, but also by dialysis technology, replacement joints and the development of sophisticated surgical technologies (i.e. micro, orthoscopic, laser, etc.). Human organs can be replaced with organs from other people or sometimes from other animals. Computer technology changes so quickly that businesses usually expect top-of-the-line technology to be obsolete in less than five years.

In contrast to the rapid advancement in "hard science technology", social technology has experienced almost no real advances in 100 years. In social science, one never finds a well-developed theory that (1) describes a phenomenon, (2) identifies its predisposing or precipitating conditions, (3) explains the mechanism through which the process works, and (4) permits the prediction and control of the phenomenon. Even Freud's famous psycho-dynamic theories fail. Although his theories distinctly describe the phenomenon and a mechanism through which predisposing factors become manifested as the phenomenon, it accounts for all unexpected observations by attributing them to "defense mechanisms", such as, repression, displacement, projection, and reaction formation. Although the theory is an excellent framework in which to understand events, it does not lend itself to verification or refutation. Theories regarding cognition, motivation, affect, and all



Thomas R. O'Neill

Thomas R. O'Neill is the psychometrician for the American Dental Association's Department of Testing Services. His professional interests include computer adaptive testing, performance assessment, cheating strategies, and the philosophical foundations of measurement. He presently has no personal life in which to have other interests...but someday...

Email: oneillt@ada.org

other important social science topics fail to adequately address these four issues. In the absence of powerful ways to verify and refute theories, researchers are left to assess the theories intuitively permitting them only to form opinions rather than empirical conclusions about the theory. As a result, every theory has proponents and opponents, which effectively thwarts any type of universal consensus. The acceptance of competing theories as all being equally good has distanced social scientists from the process of theory building. The distinction between theory building and modeling data has become blurred. If social science is to achieve the same status as physics, then the distinction must be made clear, and social scientists must shift toward building and applying theories.

A theory is a coherent set of principles that is used to explain a wide range of related observations. The quality of a theory is judged by the range of facts that it explains and the precision with which it explains them. Although explanations of past events are comforting, the value of a theory lies in the accuracy of its prediction of future events. When an observation contradicts a well-established theory, the researcher usually suspects an error in the data collection or analysis before disputing the theory because of the substantial accumulation of evidence already supporting the theory.

Modeling data is a very different enterprise. In the absence of a strong theory, researchers often collect data that they believe to be related to their topic. Assuming that truth can be found in the data, the researcher tries to find the most parsimonious mathematical representation that will recreate the observed data reasonably well. To see if these predictors will be applicable to future situations, the model must be cross-validated using a second sample. However, even when a model permits very accurate predictions across a wide variety of samples, it is still not a theory until the model can be meaningfully understood. Although models require the predictor variables to be operationally defined, their meaning may be ambiguous. Models may include variables that are correlated with the outcome, but are not conceptually part of the construct. For example, suppose that socio-economic status (SES) is moderately correlated with math ability. Although SES could be useful in making imprecise predictions about a person's math ability, it would be very difficult to coherently incorporate it into a theory of what math ability is. Variables that can't be discussed coherently in terms of the construct cannot be included in a theory. The essential difference is data modeling permits the selected observations to dominate the researcher's intentions, but in a well established theory, the researcher's intentions dominate the observations.

This difference has not always been clear to observers of physical phenomenon either. Barnett (1998) describes the historical development of the concept of time, as well as, how human needs, pre-existing concepts, and available technology influenced that concept. She provides many examples of the confusion and tension between modeling observations and applying a theory. Social scientists committed to advancing their respective fields would do well to understand how these issues have been resolved in the physical sciences. Although Barnett (1998) never addresses social science directly, the issues she highlights are quite pertinent. The following three paragraphs are a very abbreviated summary of Barnett's book with regard to some issues that are relevant to these tensions.

Primitive sundials were used to divide the day into segments, but not necessarily segments of equal size. Circa 1500 B.C. some sundials marked the calibrations for the morning and evening hours farther apart than for those hours near noon to adjust for the uneven increases in the length of the shadow cast throughout the day. This sundial produced 12 approximately equal daylight hours. However night was still a single unit of "non-day" and summer hours were longer than winter hours. Observations of the sun's position defined both the current time and the length of the hours. In the 1580s, Galileo noted that the swing of a pendulum is amazingly regular (it varies according to the length of the pendulum, not it's weight or the horizontal force applied to it). In 1657, Christiaan Huygens used this principle to build the first gravity-based pendulum clock which lost only about one second every two and a half hours. For short periods of time, this clock produced hours that were of the same duration regardless of the time of year and could work through the night. This clock produced more stable time than did observing the sun's position. Time was no longer tied to the relative position of the sun! But not entirely. In the long run these clocks tended to slow down and lose time due to friction and other factors. To rectify this, pendulum clocks had to be reset occasionally according to the only standard that was relatively stable over long periods of time, the sun and stars. The mechanical clock was not without controversy. Some people objected that it did not adequately model the position of the sun in the sky. Had the clock makers possessed the technology to accomplish such a feat, they probably would have, in effect, destroying the equal hours that they had just created. In towns, these pendulum clocks were installed in towers which permitted the town's activities to be coordinated using "local time". Methods to minimize the amount of friction in the mechanism extended the amount of time a clock could go without recalibration (resetting the time), but these improvements were limited by a precision ceiling of one second every 250 days. However, that ceiling would soon be removed.

Pierre Curie's discovery that quartz crystals vibrate at a very stable frequency when pressure (or electric current) is applied to them, led W. A. Marrison of Bell Laboratories to create the first quartz crystal clock in 1928. This clock was accurate to about one second every nine years. Today quartz crystal wristwatches are still quite popular. Despite their utility, quartz crystals are not the perfect solution. In addition to the imperfections inherently found in the crystals, the vibrations themselves cause some wear on the crystal which in turn changes the frequency with which it vibrates. Greater precision could be achieved if regularity was a property of the substance rather than form. This became possible with the new atomic theory and quantum mechanics. Atoms seem to function as a miniature solar system in which there is no friction. Using these ideas, atomic (cesium-133) clocks have been devised that are accurate to approximately one second every 10 million years.

Despite these improvements in precision, the original concepts of year and day as based upon the earth's orbit and rotation have not been vanquished. People find these models easy to understand and easy to relate to the experience of time. Although the production of stable hours, minutes, seconds, nano-seconds, etc. is better accomplished by observing more regular and controllable occurrences of nature (i.e. pendulum swings, crystal vibrations, etc.), the count of those occurrences are then incorporated back into an abstracted framework based upon the original concept. The idea of a mean solar day recognizes that the rotation of the earth is not constant. With the invention of atomic clocks that are precise to one second in 3 million years, it seemed silly to use the mean solar day as the standard from which seconds were derived. Rather than define a second as 1/86,400 (1/24x60x60) of a mean solar day, the 13th General Conference of Weights and Measures redefined a second as 9,192,631,770 oscillations between two specific energy levels of a cesium 133 atom under highly specified conditions. This, in effect, redefines a solar year as 86,400 "atomic" seconds rather than vice-versa.

The regularity of the sun's position relative to the earth's was replaced by the regularity of gravity's effect on a pendulum, which was replaced by the regularity of the vibrations of a quartz crystal, which, in turn, was replaced by the regularity of an electron's orbit around the nucleus of an atom. The discovery of finer gradations of regularity in nature permits humanity to extend the concept of time.

As these advances have occurred, the notion of time has become clearer. Time is certainly an abstraction created by man to make the world more understandable, but is the primary purpose to predict certain types of events or is it to create a framework to understand the events. When the framework and outcome agree, there is no conflict, but when there is a discrepancy, which one should dominate? If the purpose of time is to accurately predict the position of celestial bodies relative to a particular point on a rotating planet that orbits a star, then the failure of an equal interval measurement system to predict those positions indicates that adjustments should be made to the model. Furthermore, these adjustments should be made even if it degrades the interval quality of the model. This approach would be popular in pre-electrical societies whose concern is the amount of useable time (daylight) remaining before nightfall. The disadvantage is that it would be acceptable and probably necessary to have a different model for every point on the planet and for every day of the year for which you wanted a prediction. As a result, time would be very specific to location, which in turn would make coordinating operations over any distance quite imprecise.

However, if time is intended as a theoretical framework to make sense out of events, then having a stable equal interval framework is important. Rotational and orbital anomalies can then be regarded as merely imperfections in the cosmic machine rather than a shortcoming of the framework. In the quest to harness time, chronometry specialists have done two things. First, they have sought out ways to increase the regularity of the phenomenon that they observe to make their measurement system more stable. Second, they have investigated those observations that seem to depart from what the theory predicts to find why the observation was anomalous. The theory is only modified when the source of the anomaly is conceptually part on the construct of time. If the source of the anomaly is unrelated to the construct of time, then ways to remove its influence are sought out.

If social science is to experience the same rapid advancement as the physical sciences, then social scientists must improve their instruments and clarify the constructs implied by those instruments. Social scientists must free their ideas about the construct from the particular observations (modeling) and permit the theory to dominate. The lessons for social scientists are twofold. First, seek out methods that will permit finer and more stable regularities. Search for social science pendulums, quartz crystals, and cesium atoms. Second, do not attempt to incorporate the influences of extraneous forces into your theoretical framework. Control them! When creating a social science clock, seek to reduce the friction in the mechanism, control the temperature of the pendulum, and stay alert for other sources of error.

Barnett, J. E. (1998). Time's Pendulum: The quest to capture time from sundials to atomic clocks. New York: Plenum Publishing.

MEASUREMENT MUSINGS

Statistics versus Measurement?

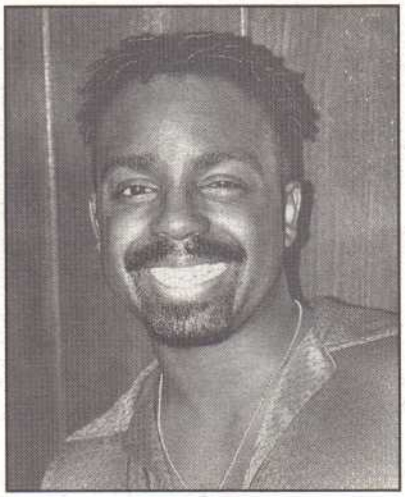
Keith M. McCoy

Most of the quantitative methods I have learned come from formal statistics. Upon satisfying all required doctoral coursework and two written prelim exams in statistics, I was stymied as to what dissertation topic to research. As a result, I embarked on a quest. How will statistics aid my career in education? As a math instructor at Chicago City Colleges, I engage frequently in testing and measuring student ability. I found what statistical theory lacked, measurement theory provided.

Parametric statistics generally involves modeling data. That is, after data collection, one seeks a model that adequately accounts for the data. This model should generally address the variability in the data. Consider this crude example. Suppose two models differ only in the amount of data variation explained by each model. A model that captures 90% variation in the data would then appear better than one that only captures 75% data variation. Sometimes a model is pre-specified. (a priori). Data often forced into a model whether they fit well or not. The idea that models and data should be independent seems lost and not investigated. When a model does not suitably fit the data, a desperate search is made for a better one. The problem may lie not with the intended model but with the data. Do the data violate the desired object of measurement? Is there a subset of the overall data that do not suitably fit the model? Is some other obscure construct being measured? These problems persist throughout educational data.

Most educators (myself included) consider themselves excellent test constructors. These are opinions not necessarily facts. Little is done to validate our tests. We regularly violate measurement assumptions by treating ordinal scores as linear measures. We assume that scores from a set of test items are additive and unidimensional. This is very far from the truth. My quest to provide better measures in testing data has led me to the school of measurement.

I certainly have a long way to go in my journey for good measures. Yet, I do not believe that the two schools, statistics and measurement, are mutually exclusive. Measurement models such as Rasch models provide researchers with appropriate linear measures. Statistical techniques like regression can be used to provide further analyses on these linear measures. As a result, I feel my journey will not be an arduous one. Moreover, many in the school of measurement are highly knowledgeable about statistics. So, I know my adventure from statistics to measurement will not be a lonely one.



Keith McCoy

Keith McCoy is an Assistant Professor of Mathematics at Wilbur Wright College. He is pursuing a doctoral degree at the University of Illinois at Chicago in Measurement. He works part-time as a reseach analyst at the American Society of Clinical Pathologists, and acts during his spare time in local Chicago Theatres.

What Works for Me

Rita K. Bode, Ph.D.
Rehabilitation Institute of Chicago

Rasch practitioners would like everyone in the world to use their model. However, if its use is to be expanded beyond its current realm, they may need to develop a variety of explanations to explain its concepts. Not everyone takes to mathematical explanations using terms such as "inverse probability" or "conjoint additivity". Some may not be mathematically inclined, others may think quantitatively but not have a taste for mathematics (a subtle difference), while still others are interested only in application. Just as an understanding of the workings of an internal combustion engine is not necessary to drive a car (as Ben Wright has said many times), understanding the mathematical foundations of objective measurement is not essential to being able to apply it. All that is required is a basic understanding of the concepts involved. There is no reason why the basic concepts could not be explained in terms with which people are already familiar, perhaps through the use of analogies. While analogies are imperfect explanations of complex ideas, they allow people to put new information into a context that they already understand. Once they get the "gist" of the idea, they can proceed to apply that idea. For some it might mean accepting the explanations that are provided without an in-depth understanding of the mathematical operations, while for others it may lead to further exploration of their foundations to truly understand them.

I'd like to call for an exchange of simple, concrete explanations of specific objective measurement concepts that work for Rasch practitioners who have had to explain them to colleagues or students. The explanations that made sense to practitioners when they learned the basics will not necessarily work for everyone. When this happens, practitioners have to develop other ways of explaining them. The explanations may only work for some people, but the greater the variety of simple explanations available, the greater the chances of finding the one explanation that will work best in a particular situation. Sharing explanations will expand the number of ways in which these basic concepts can be explained.

I'd like to start off this exchange with an explanation of misfit that has worked for me when explaining it to someone who has some knowledge of statistics. I describe the analysis of fit in terms of a chi-square analysis using the explanation provided in Chapter 4 of "Best Test Design." If someone understands chi-square analysis conceptually, they should be able to understand misfit. Is it a perfect explanation? No, but it has helped some people understand the general concepts involved in fit analysis. Here it is.

Fit analysis is a type of chi-square analysis that compares the responses observed to the response that would have been expected of the person given their responses to the set of items. Some variation from expectation will always be found because no one responds exactly as expected. But when the responses to an item or by a person exceed random variation, that variation is considered significant and evidence of misfit. Conceptually but not necessarily computationally, expected responses are determined by examining the marginal totals for a given cell. The difference between the expected and observed response is obtained and squared and these differences or residuals are summed across persons and across items. If the sum of differences across items (or persons) is not significant, the variation can be considered random and the item (or person) fits the Rasch model. But if this sum is so large as to be improbable, then the item (or person) misfits the model and is re-examined to discover why.

Rita Karwacki Bode, Ph.D.

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.

MEASUREMENT MUSINGS

Research and Practice:
Bundled Bedfellows

Robert L. Durrah, Jr.

Research and practice seem antithetical to one another. A schism exists between research professionals and teaching professionals. While researchers and teachers have much to learn from one another, often they do not find common ground for their respective endeavors until a study requires researchers to look closely into schools. Even then, researchers and teachers have little to do with each other and generally do not interact in ways that inform either group's practice (Baker & Herman, 1983; Gullickson, 1984; Rudman, 1987; Rudman et al., 1981). Since researchers, especially measurement experts, do not do much in the way of on-site school research, their literature is often obscure to school practitioners (Baker & Herman, 1983). What we have is a curious problem. Teachers do not make much use of research products in the conduct of their practice, and researchers do not discuss the implications of their research with teachers. In fact, it would seem that practitioners talk to practitioners about the craft, and researchers talk to researchers about their craft. This is a two-sided problem that seems significant. It can be thought of in the same vein as the Puritan practice of "bundling." In winter, Puritan teenage couples were over dressed and wrapped separately in tightly wound blankets. Then they were laid in a bed where they could talk and spend time together, but not touch. In the case of practitioners and researchers, there seems to be an academic "bundling," where they occupy the same bed of endeavor but enjoy no real contact.

Why is this bundling problem significant? The first answer is that both practitioners and researchers exclude critical antecedents from their work. Teachers must master multiple bodies of knowledge to be successful in their craft-general education literature, and the literature for the discipline in which they practice, and a psychological literature concerning learning. Elementary teachers have a harder job because they teach all multiple subjects to their pupils. The antecedents missing from teachers' work, are a thorough understanding of the research literature surrounding within the bodies of knowledge that frame their work in classrooms. On the other hand, the antecedents missing from researchers' work seem to be a thorough understanding of the conduct of teaching. While measurement professionals do not typically do research in classrooms (Baker & Herman, 1983), if they were to understand the context and the work that teachers do in classrooms, that understanding would go a long way toward informing researchers about better ways to measure the performance of students. While this may not be applicable to all disciplines, many of us would agree that the practitioner end of a discipline and the research end are distinct from one another. However, many researchers and practitioners will hesitate to acknowledge that there are beneficial and direct connections between research and practice.

Another reason this problem is so significant is that future knowledge advances may be delayed or lost because teachers are unable to supply students with the

Robert L. Durrah, Jr.

Born in Washington, DC and raised in North Carolina, Robert L. Durrah, Jr. did his undergraduate work at Duke University where, in 1979, he obtained an AB in History. In 1989, he completed a Masters Degree in Educational Administration and Supervision at Wichita State University, Wichita, Kansas. In his professional career he has been an Air Force Officer, Logistics, Analyst, Social Studies teacher, high school Assistant Principal, and Special Education Charter School Director.

For six of the last eight years, Mr. Durrah has worked directly with schools in varied capacities with the Center for School Improvement at the University of Chicago. At the Center, he was responsible for the implementation of social service initiatives in several Chicago elementary schools, but as the newly appointed Principal of the North Lawndale College Preparatory Charter High School, Chicago, IL, Mr. Durrah is making a transition into his new position.

Finally, Mr. Durrah's University of Chicago connection began in 1992 as a Ph D cohort member in the Department of Education. His academic interests focus on the connections between assessment and instruction at the classroom level, and how teachers use assessment to reframe their instructional inputs. Mr. Durrah will complete his doctoral studies within the year.

most current knowledge. If teachers are not privy to cutting edge technologies, or the nuances of a particular research literature, it is unlikely they will be able to introduce students to the most current knowledge available. Consequently, when students begin to pursue serious intellectual studies, they have to master greater amounts of information than they might have if they had been exposed to the most current information all along. The earlier students get current information, the more familiar they will be with particular disciplines when they begin their university careers. Rather than having to survey an entire body of information and familiarize themselves with all of it, they would be equipped to pursue new bodies of knowledge from an advanced state of acculturation and familiarity. It may be optimistic, but students so informed would be able to pursue new knowledge at the limits of what we know sooner rather than later, and they could push our knowledge beyond those limits more readily.

A caveat to this bundling problem shows up when we consider researchers who make discoveries and attempt to push the knowledge base of their discipline forward. These professionals tend to report their discoveries in research literature that is disseminated primarily amongst professionals like themselves. We do not normally recognize this as problematic, but the language of research literature tends to address the concerns of other researchers in their particularistic language. Teachers on the front lines, who could benefit directly from new knowledge, do not gain access to this new knowledge because it generally is not written for or disseminated to them.

This new knowledge could enhance teachers' work with students, and add to the knowledge base that students take with them into undergraduate institutions. Currently however, when practitioners get new information about their practice, it comes through additional university course work, in-service activities, district initiatives, or at the hands of a research-literate building administrator. One problem with accessing new information in these ways is that teachers do not always avail themselves of the opportunities for many reasons. In the case of course work, costs may be prohibitive. The information teachers can get from in-services and district initiatives can be limited or shallow. School district initiatives often require teachers to buy-in to the process or face sanctions. Sadly, while teachers may get some level of exposure to new knowledge, it is unlikely they will receive the kind of support necessary to implement the new knowledge. Finally, the opportunities teachers have to enhance their knowledge base can vary tremendously. With all the research knowledge available, the kinds of problems cited above prevent teachers from gaining access to it. Consequently, attempts to disseminate research knowledge to teachers seem about as effective as draining a water tower through a straw. Access to high quality pertinent and readable research information must be ongoing. If that access is short-circuited, the world as well as practitioners and researchers lose, and we lose unnecessarily so.

Given these problems, some readers might try to determine who is responsible. Blame fixing is inappropriate, but we do need to recognize that if we were to choose to do nothing about the problems, we would be directly responsible for them. We must focus our attention on the obvious and serious detriments to our attempts to advance knowledge. The division between practitioners and researchers hinders true collaboration between them.

Consequently, the price we pay for this disconnect between practice and research, between practitioners and researchers has been and continues to be a steep one. Because of this schism, increased preparation is required for students who would survey the breadth and depth of a research literature. That preparation very likely consumes resources that could be used to advance knowledge in the field. On the front line where teachers impart knowledge, their work is handicapped because they are unable to make use of the research that has been conducted. Because of bundling, the must crucial thing we lose is the ability to understand and expand knowledge in both the disciplines and the professional practices that rely upon those disciplines. We could abate this loss if practitioners and researchers were to work together to collaborate about both ends of the education business–practice and research.

In conclusion, practitioners and researchers are both in the same field of endeavor–education. They both are attempting to share what is known about life with the world at-large. However, knowledge disseminated only amongst the knowledgeable is of little benefit to the world at-large. How will the world benefit from research if researchers write it for themselves and share it with themselves, or practitioners discuss the practice only with one another? The answer seems simple. The world cannot benefit from knowledge in a vacuum, and knowledge in a vacuum can never be popular. The most effective research will focus on practice while the best practices will be informed by research. This is a laudable goal, one that will only be realized when practitioners and researchers achieve true collaboration; however, even a simple dialogue between practitioners and researchers would begin to bridge the great expanse of bundled knowledge that separates them.

SELECTED BIBLIOGRAPHY

Baker, E. L., & Herman, J. L. (1983). Task Structure Design: Beyond Linkage. Journal of Educational Measurement, 20(2), 149-164.

Gullickson, A. R. (1984, April). Matching teacher training with teacher needs in testing. Paper presented at the American Educational Research Association, New Orleans.

Rudman, H. C. (1987). Classroom instruction and tests: What do we really know about the link? NASSP Bulletin, 71(496), 3-22.

Rudman, H. C., Kelly, J. L., Wanous, D. S., Mehrens, W. A., Clark, C. M., & Porter, A. C. (1981). Integrating assessment with instruction: A report. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Los Angeles, CA.

Just Say "No!"

The Impact of Negation in Survey Research

Marci Morrow Enos

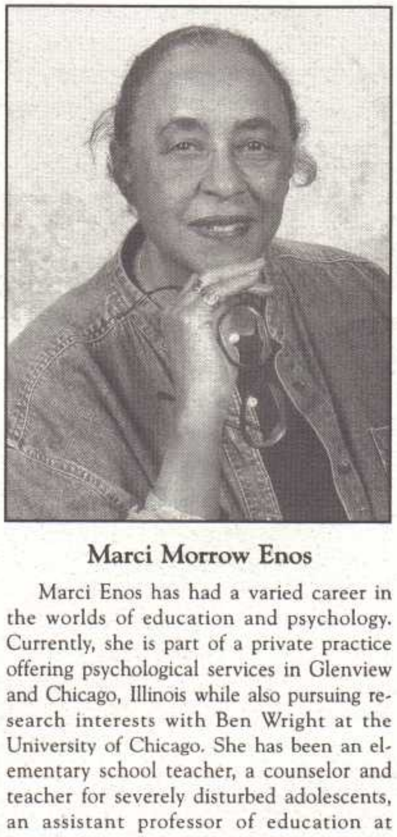
Negation may win elections, but it creates misunderstandings in survey research. Accusations of "negative campaigning" and "negative advertising" abound in political races. The implication is that candidates who use negativity take unfair advantage, since it grabs the public's attention. Negativity made headlines in the Republican presidential primary in South Carolina when Senator John McCain blamed his loss on Governor George W. Bush's "negative message of fear" (Berke, 2000, February 20).

My story is about negation's effects - not in politics, but in survey research. Long ago I was involved in the development of questionnaires to elicit students' attitudes toward school. The questionnaire items were thoughtfully chosen and closely targeted but, when the results were analyzed by Rasch methodology (Rasch, 1993/1960), a disturbing pattern emerged. The response format used four categories. Positive and negative items were included. Everything was done according to standard research methods. Negative items, which asked about the "bad" aspects of the attitudes examined, were reversed coded so that the respondent's reactions would be "in line" with their responses to the positive items. The problem emerged when the scales were analyzed with Rasch methodology. Many of the negative items misfit and were found to be measuring a variable different from the positive items.

This experience stuck with me and has led me to investigate this phenomenon. Social scientists should try to be as smart as politicians. Politicians understand the unique power of negation. Social scientists seem to think it is just affirmation flipped over!

Abuse of the Positive and Misuse of the Negative

The once popular concept of self-esteem has taken a beating in recent years. A New York Times article (Johnson, 1998, May 5) criticized a self-esteem survey instrument (Rosenberg, 1979) used in a study of educational change in the California school system. Educators and researchers expressed disappointment in the project. The results did not yield the expected correlations with aptitude and achievement and, therefore, could not predict the direction of academic progress.



Marci Morrow Enos

Marci Enos has had a varied career in the worlds of education and psychology. Currently, she is part of a private practice offering psychological services in Glenview and Chicago, Illinois while also pursuing research interests with Ben Wright at the University of Chicago. She has been an elementary school teacher, a counselor and teacher for severely disturbed adolescents, an assistant professor of education at Roosevelt University of Chicago, a test author with a publishing company and, at Michael Reese Hospital of Chicago, she was a clinical therapist and the director of a research program designed to help hearing handicapped infants and their parents. She is married and the mother of two wonderful daughters.

The whole idea of "self-esteem" was called into question.

While conceding that the self-esteem studies may have suffered "distortions in how self-esteem statistics had been gathered," the Times article cites several prominent educators who bash self-esteem as a construct:

- Research [indicates] that esteem is not in and of itself a strong predictor of success. The idea that high self-esteem is the exclusive province of those with admirable achievements has been rejected.
- Questions have been raised about the size of [self-esteem] effects and the direction of effects and whether in fact it's a mixed blessing to even have high self-esteem.
- Criminals and juvenile delinquents . . . often have high self-esteem.
- Self-esteem . . . mutated instead into a kind of crutch that explains . . . low achievement.

The baby was being pitched out with the bath water. The belief that the constellation of ideas and opinions we have about ourselves shapes how we behave makes sense. These ideas, under a variety of names – self-image, self-esteem, identity, ego, self awareness or self-concept – have long been used by human behavior researchers such as Bloom (1976), Brookover (1964), Coopersmith (1967), Epps (1969), Purkey (1970), and Rosenberg (1965). What was wrong? I reexamined the Rosenberg Scale to find out why this instrument did not lead to useful results.

Raw scores were used in the computation of esteem scores. But raw scores are not linear (Wright & Stone, 1979), and perhaps that was the problem. The inches on a yardstick are useful only because each inch is the same as the one before it and the one after. One yardstick is like another. My height is the same using my yardstick and the one at my doctor's office. Because of this uniformity, my height is predictable.

Table 1

Rosenberg Self-Esteem Scale

1. On the whole, I am satisfied with myself.
2. At times I think I am no good at all.
3. I feel that I have a number of good qualities.
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of.
6. I certainly feel useless at times.
7. I feel that I am a person of worth, at least the equal of others.
8. I wish I could have more respect for myself.
9. All in all, I am inclined to feel that I am a failure.
10. I take a positive attitude toward myself.

Scoring directions state:

Half the questions are phrased positively and half negatively. For the positively phrased questions...score as follows: Strongly Agree, 4 points; Agree, 3 points; Disagree, 2 points; Strongly Agree, 1 point. For the negative questions... reverse the scoring so that strongly agree is worth one point and so on. The maximum is thus 40 points, the minimum is 10. (NYT, 1998)

Perhaps the Rosenberg Scale (Table 1) is too abbreviated. It has only ten questions. Five of them are worded positively. This may be too few to delineate such a complex variable.

When we intend to develop a linear variable, it is important to use a range of items. The scale should include some easy items, some a bit more challenging, and some that are hard. It is unrealistic to expect only five positive items to carry the weight of self-esteem on their backs.

The Rosenberg directions say to score the negative items in the opposite direction from the positive and add them to the positive scale. Social science research has long utilized this positive plus reversed negative strategy to combat a "mind set" in the respondents. Wright and Masters (1982), citing Angell (1907), discuss this practice of constructing attitude measures from equal numbers of positive and negative statements–done with the hope of "balancing out" the effects of individual response styles. Wright and Masters show us that this strategy does not correct the problem. It is more important is to discover whether all items "provide consistent information about a person's attitude before combining them to obtain a single attitude for that person" (p. 135).

Why Isn't Negative the Opposite of Positive?

In *De Anima*, Aristotle wrote that "knowledge of the soul admittedly contributes greatly to the advance of truth in general and, above all, to our understanding of Nature" and noted further that "to attain any assured knowledge about the soul is one of the most difficult things in the world" (McKeon, 1973, p. 155).

We test designers, attempting to understand our "souls," face this difficult task when we develop survey instruments. We devise affirmative statements, targeted on our variable, which we expect respondents high in the trait will affirm. Our dream is that our respondents will treat the negative statements in a manner consistent with the way they affirm positive statements. If they "mildly agree" with a positive statement, they will "mildly disagree" with its opposite. Were this to happen, a smooth, linear variable would emerge when positives and reversed negatives are added together. Rasch analysis shows this does not happen.

This analysis reveals that to say "No" to a statement is not the equivalent of saying "Yes" to its opposite. If I should strongly reject the statement, "I hate you," it does not follow that I would strongly endorse the statement, "I love you," or even, "I like you." A negative statement is not the opposite of a positive one.

There is No "Just" in "Just Say No!"

"No" is a big deal–an

important thing to say. Ask the mother of any two-year-old. Of the many ways we try to control of our lives, an important one is our ability to refuse, to abstain, to object, to fight back, to resist, to say "No!" We don't "just" say it randomly, without some preparation, some adjustment of our mental state. Biological, developmental, linguistic, and psychological necessities are the antecedents of this behavior.

In "On Negation" (1925), Freud understands negation of a thought as a way of denying that we could have ever had that thought, thereby allowing repressed ideation to enter our consciousness. By negation, we can think about forbidden ideas.

What others might think keeps us from confessing ideas we fear would cause us shame or disapproval. We can think about forbidden ideas by denying them or by joking about them. "Thou shalt not kill," presumes our capacity for such behavior. We fear death. Yet jokingly we say, "Oh, you'll die when you hear this!" or, "I almost died when he said that!" or, "It scared me to death!"

When Less Is More: Separating Analyses

To understand negative vs. positive, I developed a longer self-esteem test from the Rosenberg items. The new test, "Thinking About Myself" (Table 2), has twenty items, ten negative and ten positive. The response format has four categories: "Strongly Disagree," "Disagree," "Agree," and "Strongly Agree."

Three forms were composed. In Form M (Table 2), the twenty negative and positive items were intermixed. In Form P, the ten positive items were given first, followed by the ten negative items. In Form N, the ten negative items were given first, followed by the ten positive items.

The forms were administered to graduate students. Some students took Form N, while others took Form P. All students took Form M, with items intermixed.

Responses from all three versions (Forms N, P, and M) were combined into one analysis. Responses were analyzed three ways: (1) responses to the 20 negative and positive items of all three forms together; (2) responses to the 10 negative items across all three forms; and (3) responses to the 10 positive items across all three forms. Because the category "Strongly Disagree" was hardly chosen for the positive items, "Strongly Disagree" and "Disagree" were combined. Responses to "Strongly Agree" and "Agree" were combined for the reversed-coded negative items.

Using the WINSTEPS computer program (Linacre, 2000) employing Rasch analysis, linear measures (logits) were constructed from raw scores. This made it possible to compare item calibrations across questionnaire versions. The analysis of combined positive and negative items yielded the "map" of items shown in Table 3.

Table 2

New Self-Esteem Questionnaire

Thinking About Myself – Form M (Mixed)

1. In general, I am satisfied with myself.
2. I think that I am no good at all.
3. I see many good qualities in myself.
4. I can accomplish things as effectively as others.
5. I am not proud of myself.
6. I feel useless much of the time.
7. I know that I am a worthwhile person.
8. I do not have much respect for myself.
9. I tend to see myself as a failure.
10. I have a very positive attitude about myself.
11. There are more successes than failures in my life.
12. It is hard for me to feel positive about myself.
13. I have a strong sense of self-respect.
14. Sometimes I just feel worthless.
15. There are many important things that I do poorly.
16. I am a useful person.
17. I am very proud of who I am.
18. My bad qualities overshadow the good ones.
19. I am dissatisfied with the person I have become.
20. I consider myself a really good person.

Table 3

WINSTEPS Map of Students' Self-Esteem Ideas

Thinking About Myself – Forms NPM

Hardest to reject :	-Just feel worthless
Hardest to agree with:	+I'm useful person
	+I do poorly
	+Have Positive Attitude
	+Very proud
Moderately hard to agree with / reject:	+Accomplish things, -Not proud
	-Have bad qualities, -Not positive about self,
Moderately easy to agree with;	+Good Person, +Strong self respect
	+Satisfied w/self,
	-Dissatisfied w/self, +More successes, +Worthwhile
Easy to reject	-A failure, -Not much self respect, -Feel useless, +Good Qualities
Very easy to reject	-I'm no good

Note: The minus in front of an item may be read as, "I'm not..." or "I reject the idea that I. . . .

Table 3 offers us a confusing story. In this WINSTEPS map, the easiest items are at the bottom, the hardest at the top. This map shows that the easiest items are negative. Respondents found it easier to reject negative items than to affirm positive items. The very hardest item was also a negative one. It was very hard to reject feeling "Worthless," although it was easy to affirm being "Worthwhile." Being worthwhile was not seen as the opposite of being worthless. Only two negative items were successfully seen as the obverse of their positives: "Satisfied - Dissatisfied" and "Very proud - Not proud" which are close on the variable line. The inclusion of negative and positive items muddles our ability to interpret this analysis.

Table 4

Positive Items Only (Measure Order)

Measure	Esteem Idea
71.7	Useful. . . . . **Hardest to Affirm**
59.2	Positive Attitude
56.9	Very Proud
51.8	Accomplish Things
48.1	Good Person
45.7	Self Respect
44.4.	Satisfied
44.4	Successes
39.3	Worthwhile
38.3	Good Qualities . . . . **Easiest to Affirm**

Table 5

Negative Items Only (Measure Order)

Measure	Esteem Idea
81.7	-Worthless. . . . . . .**Hardest to Reject**
66.7	-Do Poorly
53.5	-Not Proud
52.3	-Bad Qualities
50.9	-Not Positive
50.9	-Dissatisfied
41.1	-Useless
40.0	-No Self Respect
40.0	-A Failure
27.1	-No Good. . . . . **Easiest to Reject**

The story improves when we look at the positive and negative data side-by-side in measure order (Tables 4 & 5).

By separating them, we can discuss more lucidly what the easy and hard items are on each subscale and better understand the story the respondents are telling us. When we draw arrows between the positive items and their negative counterparts, we see differences in location on the measure line. Most egregious are "Useful - Useless," "Worthwhile - Worthless," and "Good qualities - Bad qualities." These so-called reversals evoked different reactions between positive and negative.

The Principal Components (Standardized Residual) Factor Plot and related analysis of the combined positive and negative items shows in another way how respondents reacted to the questionnaire. These two tables (Tables 6 & 7) show us how the negative items drop like stones to the bottom of the analysis. The standardized residuals of the negative items, except for "No Good," are all in the bottom half of the factor loadings, indicating once again that respondents treated negative items differently from positive.

Some students were observed to be in distress while taking Form N of the questionnaire. They complained and squirmed in their chairs. When asked later, they said the questionnaire made them feel uncomfortable by confronting them immediately with a string of negative ideas about themselves. This was an unexpected, serendipitous observation, yet in line with what we

TABLE 6

Factor Plot of Positive and Negative Items

```
Principal Components (Standardized Residual)
Factor 1 explains 3.56 of 20 variance units

      ++-----+-----+-----+-----+-----+-----+-----+-----+-----+-----++
   .7 +                              |                              +
      |                              |    A  +PosAttitude           |
      |                              |                              |
   .6 +  +Good Qualities  B          |                              +
      |                              |                              |
      |                              |            C  +Useful        |
   .5 +                              |                              +
F     |                              |                              |
A     |                              |                              |
C  .4 +                         E  D | F                            +
T     |                              |   G +Very Proud              |
O  .3 +                              |                              +
R     |                              |                              |
   .2 +                              |                              +
1     |                              |                              |
   .1 +                            HJI  +Successes,SelfRes,GoodPers +
L     |        -NoGood   j           |                              |
O  .0 +------------------------------|------------------------------+
A     |                              |                              |
D -.1 +                              |                              +
I     |                              |                              |
N -.2 +                              |                              +
G     |                              |                              |
  -.3 +                              |                              +
      |                              |i -NotProud                   |
  -.4 +                              |                              |
      |                              gh -BadQual, -NotPositive      |
      |                         bc f |      d      e -Worthless     |
  -.5 +                              |                              +
      |              -A failure  a   |                              |
      |                              |                              |
      ++-----+-----+-----+-----+-----+-----+-----+-----+-----+-----++
       0     10    20    30    40    50    60    70    80    90    100
                             ESTEEM MEASURES
```

observed to be the impact of negative stimuli. Although (Tables 8 & 9) border on the astonishing, they are more understandable in light of that revelation from the students.

Table 7

Principal Component Analysis of Positive and Negative Items

```
INPUT: ANALYZED: 47 UCSTUS, 20 SELFIDEAS, 3 CATEGORIES
-------------------------------------------------------------
   FACTOR 1 EXPLAINS 3.56 OF 20 VARIANCE UNITS
+-------------------------------------------------------------+
|      |       |         INFIT OUTFIT| ENTRY                   |
|FACTOR|LOADING|MEASURE  MNSQ MNSQ |NUMBER SELFIDEAS           |
|------+-------+-------------------+--------------------------|
|  1   |   .67 |    58.8  .71  .69 |A   10 PosAttitude         |
|  1   |   .60 |    41.2  .94  .93 |B    3 GoodQualities       |
|  1   |   .54 |    69.4 2.32 2.35 |C    6 Useful              |
|  1   |   .40 |    46.5  .83 1.57 |D    1 Satisfied           |
|  1   |   .40 |    42.1  .66  .55 |E    7 Worthwhile          |
|  1   |   .37 |    52.9 1.30 1.69 |F    4 AccomplishThings    |
|  1   |   .34 |    56.9  .46  .42 |G    5 VeryProud           |
|  1   |   .12 |    46.5  .67  .58 |H    9 Successes           |
|  1   |   .10 |    49.6 1.11 1.61 |I    2 GoodPerson          |
|  1   |   .08 |    47.5 1.03 1.00 |J    8 SelfRespect         |
|  1   |   .05 |    30.5  .78  .57 |j   12 -NoGood             |
|      |-------+-------------------+--------------------------|
|  1   |  -.55 |    41.2  .75  .64 |a   19 -AFailure           |
|  1   |  -.48 |    41.8  .82  .70 |b   16 -Useless            |
|  1   |  -.47 |    41.2  .83  .71 |c   18 -NoSelfRespect      |
|  1   |  -.47 |    61.8 1.35 1.28 |d   14 -DoPoorly           |
1  1   |  -.45 |    73.6 2.00 2.82 |e   17 -Worthless          |
|  1   |  -.44 |    46.5  .70  .62 |f   11 -Dissatisfied       |
|  1   |  -.44 |    50.7 1.04  .97 |g   13 -BadQualities       |
|  1   |  -.40 |    49.7  .84  .75 |h   20 -Not Positive       |
|  1   |  -.35 |    51.7  .42  .36 |i   15 -NotProud           |
```

TABLE 8

STUDENTS – POSITIVE ITEMS

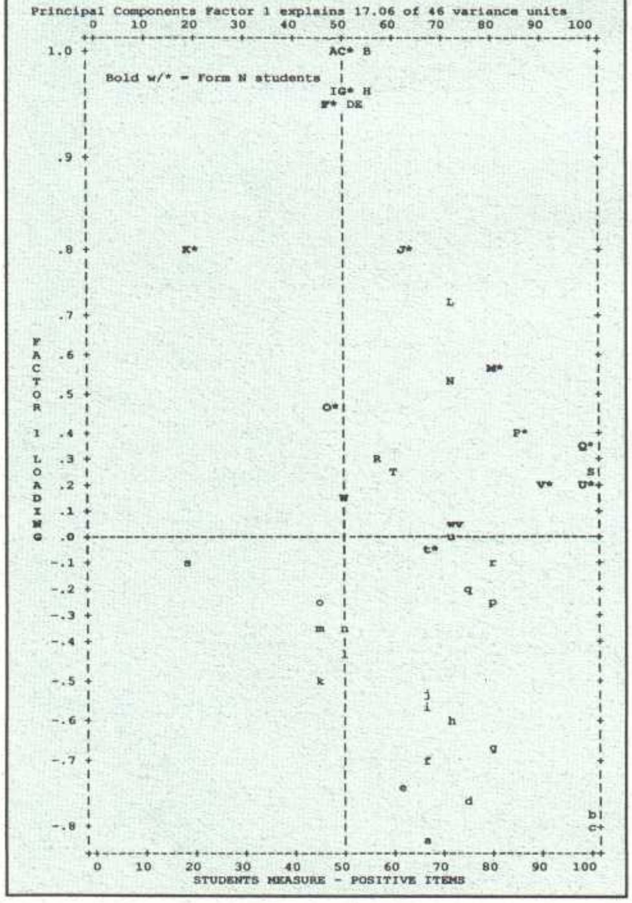


TABLE 9

STUDENTS - POSITIVE ITEMS

```
PRINCIPAL COMPONENT ANALYSIS
 STANDARDIZED RESIDUAL CORRELATIONS (SORTED BY LOADING)
  FACTOR 1 EXPLAINS 17.06 OF 46 VARIANCE UNITS
.
-------------------------------------------------------
|      |       |         INFIT OUTFIT| ENTRY       |
|FACTOR|LOADING|MEASURE  MNSQ MNSQ |NUMBER UCS    |
|------+-------+-------------------+-------------|
|  1   |   .98 |    49.9  .17  .13 |A   49 mM1   |
|  1   |   .97 |    51.6  .16  .13 |B   37 mM2   |
|  1   |   .96 |    50.4  .17  .14 |C   13 nM2*  |
|  1   |   .93 |    50.5  .16  .12 |D    1 pF1   |
|  1   |   .93 |    50.5  .16  .12 |E    5 pF2   |
|  1   |   .93 |    50.5  .16  .12 |F   17 nF2*  |
|  1   |   .93 |    50.5  .16  .12 |G   23 nF2*  |
|  1   |   .93 |    50.5  .16  .12 |H   26 mF1   |
|  1   |   .93 |    50.5  .16  .12 |I   27 mF2   |
|  1   |   .81 |    61.5 1.23 1.27 |J   11 nF2*  |
|  1   |   .80 |    19.1 1.35 2.32 |K   16 nM4*  |
|  1   |   .71 |    70.8 1.58 1.80 |L   22 nM1*  |
|  1   |   .56 |    80.1 1.22 1.26 |M   21 nM1*  |
|  1   |   .53 |    70.8 1.87 2.02 |N   32 mM4   |
|  1   |   .45 |    47.7  .07  .06 |O   24 nM1*  |
|  1   |   .39 |    85.3 1.22 1.15 |P   15 nF2*  |
|  1   |   .35 |   101.3 1.39 1.73 |Q   12 nM1*  |
|  1   |   .28 |    56.3  .41  .34 |R    4 pM2   |
|  1   |   .26 |   101.3 1.43 2.45 |S    7 pF2   |
|  1   |   .23 |    60.7 1.00 1.02 |T    6 pM1   |
|  1   |   .20 |   101.3 1.33 1.23 |U   18 nM1*  |
|  1   |   .18 |    91.8 1.23  .90 |V   20 nF1*  |
|  1   |   .15 |    50.5 2.30 2.54 |W    2 pM4   |
|  1   |   .06 |    70.8  .89  .90 |w   29 mM1   |
|  1   |   .06 |    70.8  .89  .90 |v   30 mM1   |
|      |-------+-------------------+-------------|
|  1   |  -.81 |    66.3 1.15 1.25 |a   35 mF3   |
|  1   |  -.81 |   101.3  .53  .22 |c   38 mM4   |
|  1   |  -.81 |   101.3  .53  .22 |b   39 mF1   |
|  1   |  -.76 |    75.3 1.22 1.45 |d   42 mF1   |
|  1   |  -.75 |    61.5 1.13 1.11 |e   48 mF4   |
|  1   |  -.70 |    66.3 1.32 1.40 |f   43 mF2   |
|  1   |  -.65 |    80.1 2.37 2.24 |g   41 mM1   |
|  1   |  -.60 |    70.8 2.22 2.36 |h   44 mF1   |
|  1   |  -.56 |    66.3 1.47 1.59 |i   46 mM1   |
|  1   |  -.54 |    66.3 1.37 1.48 |j   45 mM1   |
|  1   |  -.52 |    44.4  .20  .15 |k   40 mF2   |
|  1   |  -.43 |    50.5  .86  .89 |l   33 mF2   |
|  1   |  -.36 |    44.4 1.14 1.16 |m   47 mF1   |
|  1   |  -.35 |    50.5  .82  .81 |n   34 mF2   |
|  1   |  -.25 |    44.4 1.29 1.41 |o   28 mM2   |
|  1   |  -.24 |    80.1  .69  .66 |p   31 mF2   |
|  1   |  -.19 |    75.3  .74  .74 |q    3 pM1   |
|  1   |  -.12 |    80.1  .92 1.01 |r    9 pF1   |
|  1   |  -.11 |    19.1  .80  .72 |s   25 mM4   |
|  1   |  -.05 |    66.3  .81  .72 |t   14 nF3*  |
|  1   |  -.01 |    70.8  .75  .69 |u   36 mM1   |
|F,M=gender; 1,2,3,4=Age; p,n,m =version of questionnaire
+----------------------------------------------------+
```

The results yielded by the principal components analysis of the students' responses to the positive items were very interesting. Both the pictorial representation of the plot (Table 8) and the table of standardized residual correlations (Table 9) are shown. For the positive items, all except one of the students who took Form N are located in the upper (positive) region of the factor loadings (in bold, with asterisks). Note that a large portion of the variance (17.06 units) is explained by this factor.

The principal components analysis for the negative items looks very different (Table 10). On that one, the Form N students are scattered among positive and negative loadings in the expected, random way. The dramatic reaction of Form N students to the negative item bombardment was manifested mainly when they responded to the positive items. We could not have learned this if we had not analyzed the positive and negative items separately.

These analyses demonstrate the difference between

TABLE 10

STUDENTS - NEGATIVE ITEMS

```
FACTOR 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
STANDARDIZED RESIDUAL CORRELATIONS(SORTED BY
LOADING)-FACTOR 1 EXPLAINS 10.75 OF 41 VARIANCE
UNITS

+-----------------------------------------------+
|      |        |         INFIT OUTFIT| ENTRY     |
|FACTOR|LOADING|MEASURE  MNSQ MNSQ  |NUMBER UCS |
|------+-------+--------------------+-----------|
|  1   |   .93 |    44.9  .21   .17 |A   13 nM2*|
|  1   |   .93 |    44.9  .21   .17 |B   17 nF2*|
|  1   |   .93 |    44.9  .21   .17 |C   24 nM1*|
|  1   |   .93 |    44.9  .21   .17 |D   26 mF1 |
|  1   |   .93 |    44.9  .21   .17 |E   27 mF2 |
|  1   |   .93 |    44.9  .21   .17 |F   37 mM2 |
|  1   |   .88 |   105.7  .40   .14 |G   20 nF1*|
|  1   |   .88 |   105.7  .40   .14 |H   42 mF1 |
|  1   |   .69 |    56.0 1.04  1.03 |I   39 mF1 |
|  1   |   .42 |    75.3 1.11  1.14 |J   38 mM4 |
|  1   |   .35 |    65.8  .97   .92 |K   29 mM1 |
|  1   |   .35 |    65.8  .97   .92 |L   30 mM1 |
|  1   |   .30 |    61.0 1.00  1.00 |M   36 mM1 |
|  1   |   .25 |    75.3  .95   .95 |N   40 mF2 |
|  1   |   .23 |    61.0  .74   .64 |O    6 pM1 |
|  1   |   .22 |    75.3 1.10  1.12 |P   21 nM1*|
|  1   |   .20 |    94.3  .95   .77 |Q    9 pF1 |
|  1   |   .03 |    80.5 1.00   .93 |R    3 pM1 |
|      |-------+--------------------+-----------|
|  1   |  -.83 |   105.7 1.40   .53 |a   41 mM1 |
|  1   |  -.72 |    33.5 1.14  1.53 |b   28 mM2 |
|  1   |  -.62 |    28.2 3.34  8.33 |c   16 nM4*|
|  1   |  -.50 |    56.0 1.17  1.31 |d   10 pM4 |
|  1   |  -.40 |    43.8  .73   .58 |e   49 mM1 |
|  1   |  -.39 |    61.0 2.29  2.20 |f   48 mF4 |
|  1   |  -.39 |    33.5 1.14  1.10 |g    4 pM2 |
|  1   |  -.36 |    56.0  .35   .29 |h   23 nF2*|
|  1   |  -.32 |    80.5  .56   .45 |i   46 mM1 |
|  1   |  -.31 |    65.8 1.36  1.54 |j   31 mF2 |
|  1   |  -.30 |    56.0 1.35  1.35 |k   32 mM4 |
|  1   |  -.15 |    70.5  .74   .67 |l   34 mF2 |
|  1   |  -.14 |    56.0 1.55  1.74 |m    2 pM4 |
|  1   |  -.12 |    39.1 2.76  2.85 |n   25 mM4 |
|  1   |  -.12 |    80.5  .88   .85 |o   45 mM1 |
|  1   |  -.10 |    51.5  .45   .35 |p   11 nF2*|
|  1   |  -.10 |    80.5  .84   .83 |q   47 mF1 |
|  1   |  -.10 |    50.7  .42   .33 |s    1 pF1 |
|  1   |  -.10 |    50.7  .42   .33 |r   35 mF3 |
|  1   |  -.10 |    65.8  .76   .71 |t   22 nM1*|
|  1   |  -.08 |    70.5 1.79  2.00 |U   15 nF2*|
|  1   |  -.04 |    65.8  .80   .82 |T   33 mF2 |
|  1   |  -.04 |    75.3 1.03   .94 |S   14 nF3*|
+-----------------------------------------------+
```

how we react to positive and negative statements. Remember, these are just sentences on a piece of paper, no curse words, no punches were thrown, no mud was slung - or so we thought.

Although a small study, this analysis revealed a surprising trend. The map of mixed items along the logit "ruler" is muddled by the inclusion of both negative and positive items. The "story" about what is easier and harder to believe about ourselves is immediately clearer when negative and positive items are analyzed separately. When we look at the principal components analysis of the mixed items, we see the negative items showing they are a separate factor.

The principal components analyses give us evidence that the students who took Form N, with all the negative items first, experienced a common reaction. What was it? Anger? Anxiety? Depression? Whatever it was altered their behavior when they took the positive items both at the end of their Form N questionnaires and also mixed on the Form M questionnaire. For a few moments, Form N students were more like one another than they were before they undertook this task. This hints at the impact of other kinds of more active, traumatic negative experience, particularly educational evaluations.

The "Moral of the Story" is that negation is not the opposite of affirmation. Negativity has a powerful effect. Negative and positive items are not additive. This study brings out the inherent and previously unacknowledged confusion that occurs when we use an arithmetical maneuver to solve what is actually a profound psychological misunderstanding.

The usefulness of the idea of self-esteem (and many other "self" ideas examined over the years, such as motivation, aspiration, and sense of control) might not be at an end after all. What needs to be abandoned is the way survey instruments which attempt to target these variables are analyzed. Thoughtful analysis using Rasch methodology to construct useful measures from responses will make it possible to construct stable inferences from these old friends.

References

Angell, F. (1907). On judgments of "like" in discrimination experiments. American Journal of Psychology, *18*, 253-260.

Berke, R. L. (2000, February 20). Bush halts McCain in South Carolina by drawing a huge republican vote. *The New York Times*, p. 1.

Bloom, B. S. (1976). *Human characteristics and school learning.* New York: McGraw-Hill.

Brookover, W. B., & Thomas, S. (1964). Self-concept of ability and academic achievement. Sociology of Education, 37, 271-278.

Coopersmith, S. (1967). *The antecedents of self-esteem.* San Francisco: W. L. Freeman.

Epps, E. (1969). Correlates of achievement among northern and southern urban Negro students. Journal of Social Issues, 25, 55-71.

Freud, S. (1959). Negation. In L. Strachey (Ed. and Trans.) *Sigmund Freud: Collected Papers. Vol. 5,* (pp. 181-185). New York: Basic Books. (Original work published 1925)

Johnson, K. (1998, May 5). Self-image is suffering from lack of esteem. *The New York Times*, p. B12.

Linacre, J. M. (2000). WINSTEPS (Version 2.98) [Computer software]. Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. (1998). *A user's guide to BIGSTEPS WINSTEPS: Rasch-model computer programs.* Chicago: MESA Press.

McKeon, R. P. (1973). *Introduction to Aristotle* (2nd ed.) (pp. 153-245). Chicago: University of Chicago Press.

Purkey, W. W. (1970). *The self and academic achievement.* Englewood Cliffs, NJ: Prentice-Hall.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press. (Original work published 1960)

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rosenberg, M. (1979). *Conceiving the self.* New York: Basic Books.

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

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

A Standard Vision

Gregory E. Stone, Ph.D.

Who passes and who fails? What does it mean to pass? How can a fair and meaningful standard be established? Such questions are routinely asked within many different educational and evaluative settings. The stakes are high, the requirements important – a public at large depends upon these measurement devices to graduate and pass qualified candidates.

There are as many different models, empirical and otherwise, for establishing passing standards as there are examinations themselves. Some reflect complex relationships between statistical technique and judgement making, others a simplicity of qualitative purpose. All attempt to create a reasonable decision, and most are subject to significant criticism on grounds of equity, precision, and meaningfulness. In this article a conceptual and fundamental framework within which all models may be evaluated is discussed.

Regardless of the model, every standard setting method must effectively demonstrate the desired criterion, be reproduceable, and remain genuine. It is important to note that in the efforts of standard setting, golden rods and sacred cows are of little use. Ultimately the process is genuinely evaluative, and it becomes the goal of the standard setter to define a systematic, logical and understandable quantifiable method for conduct of this qualitative exercise.

The first requirement, effective demonstration of the desired criterion, is fundamental. In criterion referenced standard setting, the criterion hopes to represent a specific body of content knowledge. Theoretically, the act of passing a test demonstrates successful mastery of this content. This interpretation of a passing outcome is only reasonable if the standard adequately reflects the content. A demonstration of adherence to content I propose to call criterion validity, in support of the criterion referenced standard. While a departure from common quantitative descriptions of validity of criterion standards, it appears both logical and desireable. Unfortunately such validity is achieved very rarely.

Traditional standard setting systems (like Angoff, for example) gather together groups of experts in a subject area and ask them to predict candidate performance. A typical question posed to these experts is "how many examinees out of 100 will answer each item correctly?" Summations and averages of these predictions of performance ultimately become the standard.

Even a superficial review of such a judgement making process reflects that the desired content-based criterion is being missed. Outcomes are necessarily linked to data input. When predictions of performance are used as 'input' it follows that the products of that predicted performance becomes the 'output'. The criterion emerging from predicted performance must be a performance criterion, not a content criterion.

To establish a content-based standard, judges must define the criterion in a manner that addresses it directly. Meaningful definition is only achievable through an exercise focussing on a qualitative evaluation of the concepts within the subject matter, rather than via unwarranted and impractical predicated quantities. Thus far, only Rasch-based models have been able to demonstrate effective content validity. In particular, the Objective model (Stone, 1994, and Gross and Wright, 1965) collects judgements in terms of essentialness of content presentation, and has successfully demonstrated a singularity between qualitative judgement and quantitative outcome. Objective models allow content experts to be content experts – by selecting content of importance.

The second quality, that of reproduceability, is a concept not foreign to measurement. Generally considered reliability in quantitative circles, it is a question of reproduction of results. Standards must be able to demonstrate that they are applicable on more than a single version of an examination. A criterion 'standard' implies a level of achievement within a criterion. If the standard changes with each unique examination or grouping of items, how can a reasonable level of achievement be considered? A simple test of reproduceability is available to check standards.

Consider the passing rates for two content-similar, but not necessarily item-identical, examinations. If the standard is reliable, should the passing rates not also be the same? Not necessarily. There are three facets in a typical examination setting – the difficulty of the particular examination, the abilities of the examinees, and the standard used for passing. Theoretically the first two vary, whereas the latter (the standard) should not. To test for reproduceability, the examination forms must first be equated (in Rasch methodology most likely through common-item equating). Using a standard linear transformation, differences in examinee ability between the two groups can be controlled. The result will be two different groups of examinees where difficulty and ability are controlled. Testing for reproduceability (consistency) is as simple as visually inspecting the pass rates for each group. If identical (within the defined error), then the standard defined meets this requirement for reproduceability – and is, in short, reliable.

The third quality of a useful standard finds its roots in genuine scientific credibility. In few other aspects of measurement has this been such a pervasive problem. Unfortunately standards and standard setting is such a politically sensitive issue that the methods themselves have tried to adapt to these number games. Is 60% too low a pass rate? Then move the standard up to a level that will pass 70%. Don't call it fudging, call it "adjusting" and try to find a statistic (maybe the SEM or Mean person performance) that can somehow be used to justify the move. Standard setting is notorious for fudging.

In the real world, political and other considerations are important and often impact upon measured, considered decisions, like standards. Apart from politics, the real issue for the measurement professional is one of honest reflection. When standards must be changed, the role of a measurement expert is to express those changes and educate the stakeholders. What sort of content knowledge is being left out of the new standard? How may curricula be informed to raise the level of student performance? Instead of addressing these changes directly, many choose complicated "adjustment" techniques and errantly believe that the standard has somehow remained the same, just adjusted or corrected. Research honesty and integrity in creating a genuine standard that remains true to its defined meaning is imperative for the process.

Ultimately there may be many ways to define performance standards. However, there are at least three fundamental qualities that may be used to judge their merit. The redefined notions of validity, reliability and genuineness should be considered performance benchmarks. While only one model has thus far demonstrated each – the Rasch-based Objective model - the article expresses a desire that other models too will put themselves to these simple, yet fundamentally necessary tests.

To illustrate one way, through which the reliability of passing standards may be assessed, consider Figures 1 and 2. Each presents data concerning the passing rates observed on four national, high-stakes examinations. Each uniquely created exam was constructed using the identical content outline, but each contained a different set of specific items. The diamond pointed line represents actual passing rates on each successive administration using the same (equated) standards. The square pointed line represents what the passing rate would have been had difficulty of the examination and group person ability been controlled. A glance at the figures shows a clear linearity within the Objective standard - evidence of its reliability - while the Angoff standard does not. Instead, the Angoff standard itself or the error associated with it, produces wildly different results from administration to administration. Such results suggest a fairly unreliable process. Which passing rate should one believe? Why the fluctuation when all moveable factors have been controlled?

Figure 1

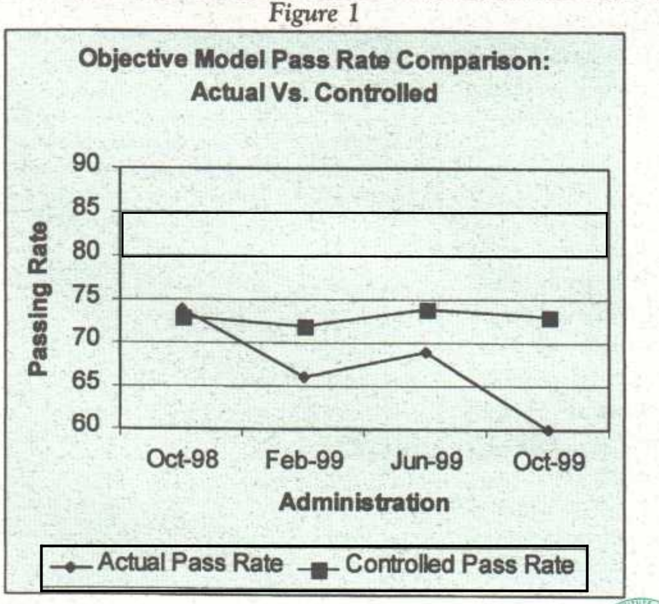
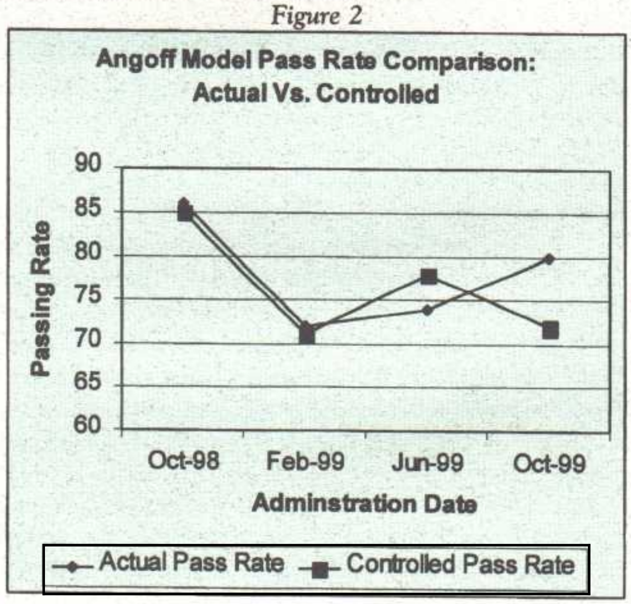


Figure 2



Precision Versus Practicality

Cindy Brito, MPH, MT(ASCP)

Histology technologists play an important role in the pathology laboratory. They are responsible for the handling of surgical tis sue speci-mens which must be processed, embedded in paraffin blocks, sliced into thin segments, placed on a glass slide, stained, coverslipped and labeled before presentation to the pathologist. The pathologist can then microscopically examine the tissue on the slide and render a diagnosis of healthy or diseased.

The American Society of Clinical Pathologists (ASCP) has, for many years, administered a practical examination in histology. After completing an appropriate course of study, a qualifying candidate submits fifteen stained slides and tissue blocks of varying difficulty which are scored by a panel of judges on a semi-annual basis. A candidate whose slides are of a quality deemed acceptable by the judges and who also successfully passes a written examination is awarded a certificate, and is eligible to Place the initials HT(ASCP) following their name. This professional designation is nationally recognized as the Gold Standard for technologists working in the histology laboratory.

The judging process begins with the selection of 20-25 pathologists and histology techs from across the nation who are asked to volunteer their time for the important project. All judges are flown to Chicago where a marathon 2-1/2 day grading session takes place. Using well-defined guidelines and standards, blocks and slides are reviewed and graded using either dichotomous (1 = acceptable and 0 = unacceptable) or 4 step rating scales (3 = excellent, 2 = acceptable, 1 = marginal, 0 = unacceptable). The results are then analyzed using a Rasch multi-faceted model (John Linacre's Facets).

The histology practical grading session has traditionally been subsidized by the ASCP. In essence, a candidate seeking certification pays the same fee as candidates do for other certification exams not including a practical. While the judges volunteer their time, the ASCP assumes all expenses for airfare, lodging, and meals for the approximate 25 judges required. The estimated cost to grade each practical is $400. In an effort to be more fiscally controlled without increasing the financial burden to the candidate, a study was undertaken to determine if the resources required to grade a practical could be streamlined, i.e., use fewer judges in each grading session. The time required for a judge to grade a set of slides has been well established over the years. Therefore to reduce the number of judges required, the choices were two: increase the number of days in the grading session, or

Cindy Marie Brito, MPA, MT(ASCP)SC

Cindy Brito is a Medical Technologist and works at the American Society of Clinical Pathologists as a Manager of Research and Evaluation in the Board of Registry department. Interests include camping, Hosta gardening and reading Ben Wright's "Best Test Design" to her cat Tigger.

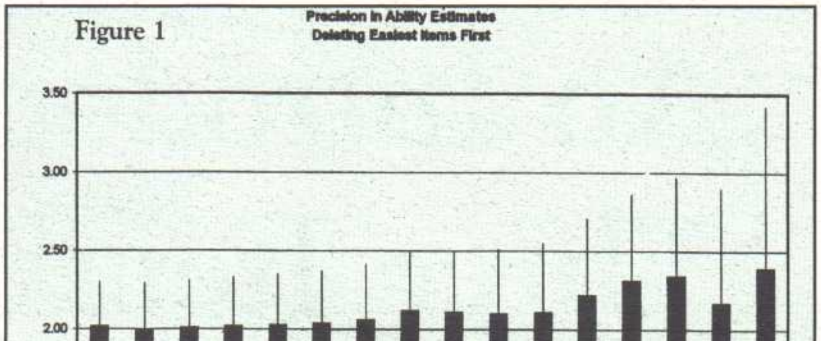
decrease the number of slides being graded.

The grading session takes place over a weekend and is a very demanding full two-day schedule for the volunteer judges. After much consideration, it was decided that another day of judging would be mentally exhausting and a fatigue factor could set in. Thus an analysis of the data was conducted to determine if reducing the number of slides would yield results that were psychometrically equivalent to the fifteen slide/block practical.

The slides each candidate submits are equally divided into three groups. There is a random assignment of the groups to judges and most practicals have input from three different judges. Another judge grades the qualities of coverslipping and block characteristics.

Using data from the May 1999 grading session, a range of scenarios were evaluated and compared to the baseline conditions described above. Eliminating the coverslipping and block scores had negligible impact on both the mean ability and precision of the scores. Next, slides were "peeled" away one by one starting with the easiest. As can be seen by the data in Figure 1, the mean ability remains stable across nearly the entire range of slide deletions until the level of two slides is reached. A decision was made that any decrease in the number of slides must be made in multiplies of three in order to maintain the judging system in place. Table 1 summarizes the mean precision and the numbers of candidates who pass and fail with each three-slide decrease. Note that some precision in the score is lost and the pass rate decreases slightly as slides are eliminated.

Figure 1



The final question weighs precision and pass rate against finances. With each three-slide decrease, the number of judges required is reduced by approximately twenty percent, which reduces expenses by 30%. The committee reviewing the data struck a balance at nine slides. At this level, the mean ability of the candidates remains the same, the precision changes by 0.09 logits, and the pass rate decreases by 6%.

Implementation of the reduced slide practical will be effective starting with the year 2000. It is a win-win situation. Candidates will not be charged a fee for the practical portion of their exam, results are psychometrically valid and comparable to the fifteen-slide exercise, and the ASCP gets to shave $125,000 off of their operating budget for the year!

Table 1

Practical May 1999

	15 slides 9 blocks 1 coverslip	15 slides	12 slides	9 slides	6 slides
Pass	127	127	120	118	113
Fail	18	18	25	27	32
Mean precision of score	+/- 0.28 logits	+/- 0.29 logits	+/- 0.32 logits	+/- 0.37 logits	+/- 0.44 logits

An Introduction to Three Item Testing

Kirk Becker
The Riverside Publishing Company

Using the Rasch model, the characteristics of a test or survey can be examined despite the presence of missing data, but is this also true about the characteristics of a population? In other words, is it always necessary to administer a test or survey in full in order to find out about a population of interest?

In order to compare the means of two populations on an instrument, many would say that all items on the instrument must be administered. Although this might be true for a completely untried test or survey, once the items have been calibrated only three items are needed. When items have been scaled using a population as a reference point, this reference point (the difficulty of the items in logits) can then be used to measure the ability level of individuals and the mean ability level of groups, in the same units. The Rasch model allows for a direct transformation between raw scores and logit measures. If a population mean in logits is known relative to a set of item calibrations, the population mean in raw score units can then be determined. For studies in which the population parameters are the main point of interest, this can mean huge savings in terms of time and money.

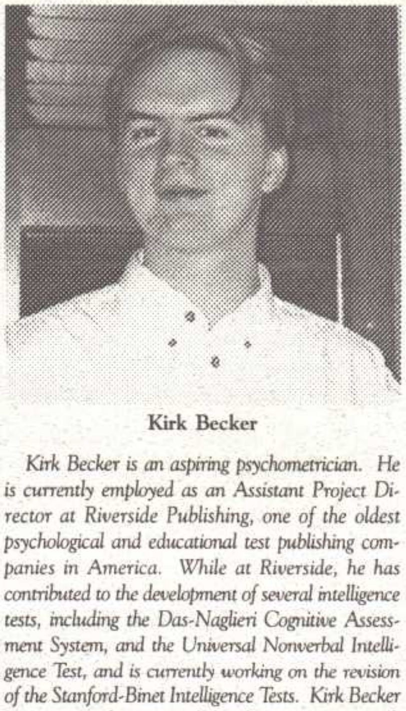
How is it possible to estimate population parameters without administering a complete measure to a large, representative sample? Data collected during the development of the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCullum, 1997) and the Stanford-Binet Intelligence Test: Fourth Edition (Thorndike, Hagen, & Sattler, 1996) was used to investigate this question.

Any pair of variables contains a great deal of information about a population that answers them. Consider the performance of 9-year-olds on a pair of items from the UNIT:

Table 1

		Item 19	
		Right: 1	Wrong: 0
Item 16	Right: 1	$178=S_{11}$	$35=S_{10}$
	Wrong: 0	$76=S_{01}$	$68=S_{00}$

If most individuals in a population fail the pair of items (S_{00}), then the population mean should logically be lower than the difficulty of the two items. Likewise, if the majority of a population pass a pair of items (S_{11}), then the population mean should logically be higher then the difficulty of the items. The ratio of S_{11} to S_{00} is therefore related to the mean of the population on the entire test, however it is also



Kirk Becker

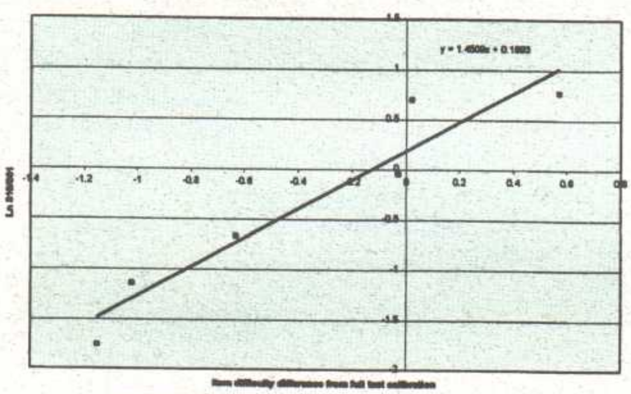
Kirk Becker is an aspiring psychometrician. He is currently employed as an Assistant Project Director at Riverside Publishing, one of the oldest psychological and educational test publishing companies in America. While at Riverside, he has contributed to the development of several intelligence tests, including the Das-Naglieri Cognitive Assessment System, and the Universal Nonverbal Intelligence Test, and is currently working on the revision of the Stanford-Binet Intelligence Tests. Kirk Becker has continued to research psychometric issues with the aid of Dr. Benjamin D. Wright at the University of Chicago. He obtained his Bachelor of Arts degree in 1995 from the University of Chicago.

a function of the item difficulty difference. The other two cells in the cross-tabulation (Table 1) are highly related to the difference in difficulty between the items. If item 19 had been very easy and item 16 very difficult, most of the population would have fallen into cell S_{01}. Likewise, if item 19 were difficult and item 16 easy, most of the population would have been in cell S_{10}. In order to examine how these relate to item difficulty and population mean, the following ratios will be.used:

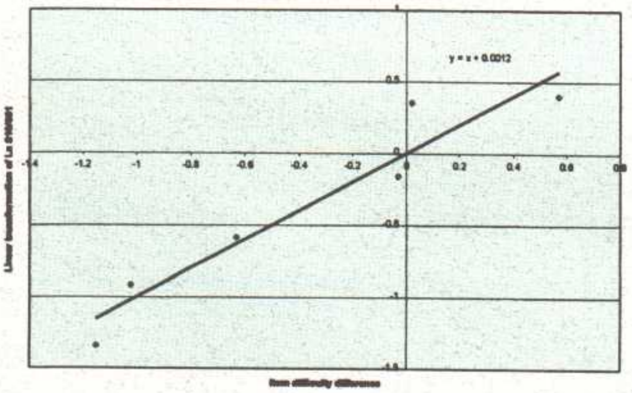
Log (S_{11}/S_{00}) Log (S_{10}/S_{01})

To examine the effect item difficulty difference has on the first relationship, the cross-tabs of several item pairs were examined. For cross-tabs between one item (item 19) and a set of other items, log (S_{11}/S_{00}) and log (S_{10}/S_{01}) are both directly related to the difference in difficulty between the items. Conceptually, the ratio log (S_{10}/S_{01}) should reveal the difference in item difficulty for a pair of items, and as Graph 1 shows, this relationship is born out. Because the mean item difficulty is set to 0, the scale of the item calibrations differs from that of the ratio, however a simple linear transformation allows us to place these sets of values on an identity line (Graph 2).

Graph 1. Item difficulty difference versus Ln S10/S01

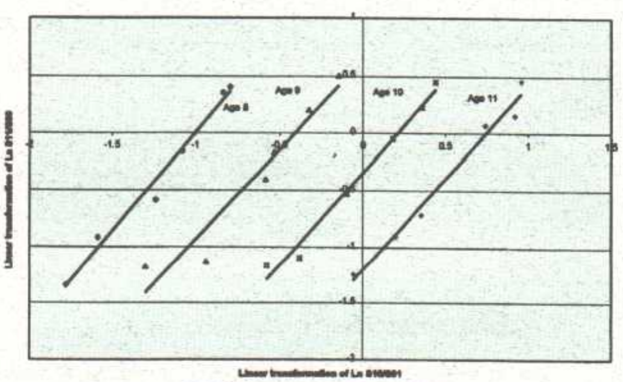


Graph 2. Item difficulty difference versus linear transformation of Ln S10/S01

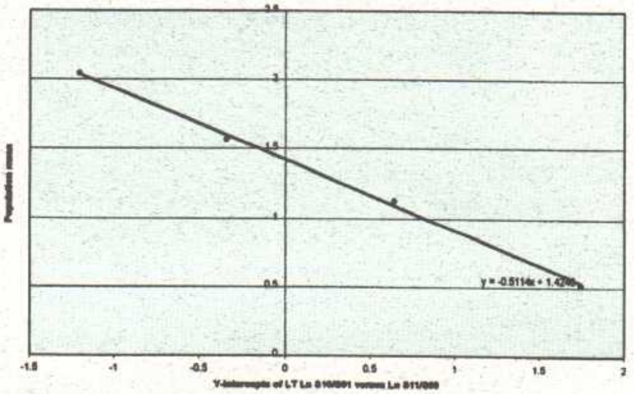


This same linear transformation can then be applied to the other ratio, log (S_{11}/S_{00}), so that both units of measurement are comparable. Once this is done, the plot of log (S_{10}/S_{01}) against log (S_{11}/S_{00}) provides a y-intercept which is directly related to the population mean. Graph 3 shows these plots for several different populations, while Graph 4 shows how the y-intercepts are related to the population means.

Graph 3. Linear transformations of Ln S10/S01 versus Ln S11/S00 for several populations



Graph 4. Y-intercepts versus population means



The formula for scaling the y-intercept of log (S_{10}/S_{01}) versus log (S_{11}/S_{00}) to the population mean is known in this case because the means are known. The slope of this line appears to be constant (m=-0.5) across multiple tests and populations. As Graph 5 shows, the intercept is the difficulty of the constant item in the cross-tabs.

UNIT
Analogic Reasoning subtest: population mean = –.51x + 1.4

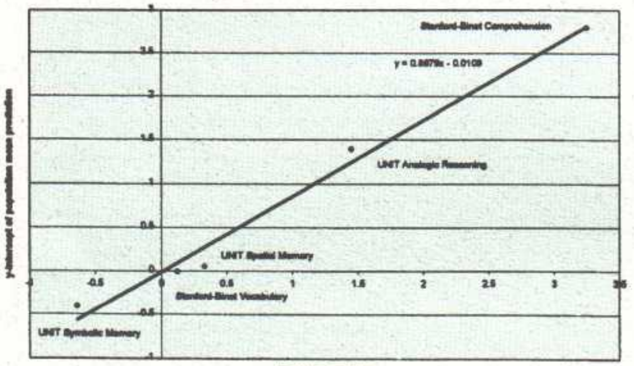
Symbolic Memory subtest:
population mean = –.4x - .41

Spatial Memory subtest:
population mean = –.4x + .06
Stanford-Binet

Vocabulary subtest:
population mean = –.42x

Comprehension subtest:
population mean = –.54x + 2.8

Graph 5. Predicting y-intercept of mean prediction from fixed item difficulty



To summarize, the steps for estimating a population mean from 3 items are as follows:

1. Administer three items from a test that has been calibrated.
2. For the two pairs of items (AB and AC) calculate the ratios log (S11/S00) and log (S10/S01) for the population of interest.
3. Perform a linear transformation on log (S10/S01) so that the plot of log (S10/S01) versus A-B and A-C is an identity.
4. Using the same scaling factor, perform the same linear transformation on the two log (S11/S00) values.
5. Determine the y-intercept of the rescaled log (S11/ S00) versus log (S10/ S01) plot for the two item pairs.
6. The y-intercept should be related to the population mean according to the following formula

Population mean = -1/2 * (y-intercept) + (difficulty of A)

References

Bracken, B. A., & McCallum, R. S. (1998). The Universal Nonverbal Intelligence Test. Itasca, IL: Riverside.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). The Stanford-Binet Intelligence Scale-Fourth edition: Technical manual. Itasca, IL: Riverside.

A Review of CAT Review

Renata Sekula-Wacura

The American Society of Clinical Pathologists administers 20 fixed length (100 item) registry examinations for laboratory professionals. Until 1993, the testing was of the paper and pencil variety, and a candidate was free to review items and change answers up to the time limit of the test. A computer adaptive test (CAT) administration was adopted in 1994. During a CAT, each examinee is administered a unique 100 item test (selected from an item bank of 500+ items) that is tailored specifically to their ability. Each item in the item bank has been calibrated for difficulty using a Rasch model (Wright and Stone, 1979). The candidate is first presented with an item whose calibration value is near or at the pass cutoff point for that exam. If the item is answered correctly, the computer program next presents a more difficult item. If the item is answered incorrectly, an easier item is presented, and so on.

The ASCP CAT program incorporates a review session. During the computer adaptive portion of the test, a candidate is required to answer all 100 items in the order presented. During this portion, any item can be marked for later review. After completing all 100 items, the computer adaptive portion is over and the program shifts into a review session. During this session, the candidate is free to look at any question in any order and to change answers until the time limit of the test is reached.

What effect does the review session have on the final score (person ability measure) and pass/fail decision? To answer this, ability measures pre and post review were examined for a period of three years. Table 1 summarizes the data. Out of 29,293 candidates, 67% passed before review and 69% after review. Table 2 summarizes the effect of the review session on the pass/fail decision. From Table 2 it can be determined that 1300 candidates had their decision altered due to the review session (pass to fail, or fail to pass). The good news is that three times as many candidates who changed answers improved their scores by doing so as opposed to those that lower their scores.

Table 1. Summary of test outcome before and after review.

BEFORE REVIEW PASS	BEFORE REVIEW FAIL
19,517 67%	9,776 33%
AFTER REVIEW PASS	AFTER REVIEW FAIL
20,197 69%	9,096 31%

Table 2. Sumary of test pass/fail outcome before and after review

PASS TO PASS	PASS TO FAIL
19,207 66%	310 1%
FAIL TO PASS	FAIL TO FAIL
990 3%	8,786 30%

What are the candidates doing in the review session? Are they changing many answers or just a few? To answer this a table was created based on all candidates. A difference in the candidate pre and post review measure and the deviation of their difference (based on the candidate standard error) was calculated.

Figure 1. Candidate measure map

```
                                                          Pre-Review


                                                             P
                                                             A
                           -3.0      -2.0      -1.0      0   S    1.0      2.0       3.0
Item Ans √ =  Time !       |++++|++++|++++|++++|++++|++++|+++S|++++|++++|++++|++++|++++|
   1  1  0      0'34                                         |+
   2  1  0      0' 3                                         +
   3  1  1      0' 2                |                   X    |+            |
   4  1  0      0' 2             |                 X         +       |
   5  1  0      0' 1            |               X            +    |
   6  1  0      0' 1          <               X             +| |
   7  1  0      0' 1          <              X               |+
   8  1  0      0' 1          <             X               ||+
   9  1  0      0' 2          <           X                |  +
  10  1  1      0' 5                |           X           +|
  11  1  0      0' 2               |            X          | +
  12  1  1      0' 2                     |         X         |+
  13  1  0      0' 1                    |         X         |+
  14  1  0      0' 1                   |         X         | +
  15  1  0      0' 2                   |        X         |  |+
  16  1  0      0' 0                  |        X          |  |+
  17  1  0      0' 0                 |         X         |   +
  18  1  0      0' 0                 |        X          |   |+
  19  1  0      0' 2                |        X          |   +|
  20  1  0      0' 0                |        X         |     |+



                                                                Review


                                                            P
                                                            A
                          -3.0      -2.0      -1.0      0   S    1.0      2.0       3.0
Item  Ans √  =   Time !   |++++|++++|++++|++++|++++|++++|+++S|++++|++++|++++|++++|++++|
   1   1  0  o   0'53                                       |+
   2   4  0  =   1'26                                       +
   3   1  1  o   1'25              |                   X    |+            |
   4   2  1  +   1'41                         |             *                |
   5   3  0  =   1'43                      |             X  +          |
   6   2  1  +   0'50                           |          +X             |
   7   3  1  +   0' 9                              |        |+ X            |
   8   2  1  +   1' 4                                 |     |+   X           |
   9   4  1  +   1'25                                   |   +      X           |
  10   1  1  o   0'16                                    | +|        X          |
  11   4  1  +   0'55                                      |+         X          |
  12   1  1  o   0'53                                       |+         X          |
  13   2  0  =   2'42                                     | +       X         |
  14   3  0  =   0'44                                    |  +     X        |
  15   2  1  +   1'34                                     | |+     X        |
  16   3  1  +   0'58                                      ||+      X        |
  17   3  1  +   0'20                                       +        X        |
  18   2  1  +   1'12                                       |+        X       |
  19   3  1  +   1'41                                      +| |       X        |
  20   2  0  =   3'12                                       |+       X       |
```

X is candidate's ability

+ is item difficulty level

√ column indicates if the question is answered right (1) or wrong (0)

= column shows how question was changed: + from wrong to right, = from wrong to wrong, - from right to wrong

Ans column indicates answer candidate choose

| indicates standard error limits

Table 3. Candidates who chanced more than 25 questions with deviation from their standard error greater than 2.

Deviation from standard error (in logits)	Difference between pre and post review measure	Number of questions changed	Final pass/fail outcome
2.09	0.45	28	CHANGED
2.09	0.45	28	NOT CHANGED
2.09	0.47	29	CHAŃGED
2.10	0.47	26	NOT CHANGED
2.10	0.46	29	NOT CHANGED
2.13	0.52	36	NOT CHANGED
2.13	0.45	96	CHANGED
2.14	0.55	43	NOT CHANGED
2.16	0.52	44	NOT CHANGED
2.16	0.53	27	NOT CHANGED
2.30	0.50	31	CHANGED
2.32	0.53	32	NOT CHANGED
2.49	0.55	55	NOT CHANGED
2.52	0.58	30	CHANGED
2.54	0.54	27	CHANGED
2.55	0.61	26	NOT CHANGED
2.57	0.54	37	CHANGED
2.70	0.59	38	CHANGED
2.84	0.60	38	NOT CHANGED
3.12	0.76	42	CHANGED
3.16	0.67	28	NOT CHANGED
3.28	0.69	30	NOT CHANGED
3.36	0.72	47	NOT CHANGED
3.47	0.73	45	CHANGED
3.51	0.86	40	CHANGED
4.08	1.10	26	CHANGED
4.20	0.89	49	NOT CHANGED
4.32	0.91	72	NOT CHANGED
4.92	1.04	67	NOT CHANGED
5.58	1.19	88	NOT CHANGED
6.08	1.31	58	CHANGED
6.95	1.49	81	CHANGED
7.13	1.73	54	CHANGED
7.84	1.68	87	CHANGED
12.38	3.33	89	CHANGED

Of the 29,293 candidates, 99% changed 25 or fewer answers. Of the 1% who changed more than 25 answers, 88% had a deviation of their difference in their measure equal to or less than 2 standard errors. The candidates with a difference of greater than 2 standard errors are summarized in Table 3.

Of interest are candidates with deviations greater than 4 logits and changing more than 50% of their answers. The CAT program can generate a Candidate Measure Map that provides information about both the computer adaptive (pre-review) and review session portions of the test. Several maps were printed and a "cheater" strategy was detected. Figure 1 shows the first twenty items of both the pre-review and review sessions for one of the candidates. In the pre-review session, the candidate selected the answer "1" for every item. The time column indicates that sufficient time did not elapse for the item and distracters to be read before proceeding to the next item! The review session is where the candidate actually "took" the test. Items were read and appropriate answers were selected to the best of their ability.

The presumed purpose of the "cheater" strategy is an attempt to get an easier test. The CAT algorithm can detect this. After 40% of the answers are incorrect, the program will automatically select items with a measure close to the pass point. However, a very able candidate will likely get a test with an average difficulty below their ability.

This review of the CAT review has raised some interesting topics for further research. For example, can incorporating a minimum time requirement before allowing presentation of the next question eliminate the "cheater" strategy? Is there a correlation between the pass/fail decision and the number of answers changed? Is the candidate's true ability underestimated when the "cheater" strategy is employed? And finally, does the cut-point level of the exam influence the percentage of candidates going from pass to fail or fail to pass?

Stay tuned!

Renata Sekula-Wacura, MS, is Manager of Database and Network Operations at the ASCP Board of Registry. She enjoys relaxing on the beach and climbing mountains in her free time.

Factors that Impact Analytic Skill Ratings

Jessica Heineman-Pieper
Mary E. Lunz
Measurement Research Associates, Inc.

Jessica Heineman-Pieper

Jessica Heineman-Pieper is working towards a Ph.D. in Psychology and The Conceptual Foundations of Science at the University of Chicago. She is also training to be a licenced clinical psychologist, and is a research associate at Measurement Research Associates. Jessica majored in American History and Literature in college and graduated Phi Beta Kappa (junior year) and with high honors from Harvard University. She was subsequently awarded a Rhodes Scholarship and graduated with honors in Philosophy and Psychology from Oxford University. Jessica hopes to teach and practice the new psychology Intrapsychic Humanism, popularly articulated in the parenting book "Smart Love: The Compassionate Alternative to Discipline That Will Make You a Better Parent and Your Child a Better Person."

Mary E. Lunz

Mary E. Lunz earned a Ph.D. from Northwestern University. After teaching and consulting for several years, Mary worked as Director and Psychometrician for the Board of Registry of the American Society of Clinical Pathologists for 17 years. During this time, she began working with Ben Wright and Michael Linacre on issues relating to performance examinations, and computerized adaptive testing. Research is still ongoing. Mary is currently Director and Senior Associate at Measurement Research Associates, Inc.

Holistic ratings lack sufficient information to measure candidates with the accuracy required for high-stakes certification examinations. When examiners make only one holistic rating of candidate performance, decisions about candidate ability are consumed with measurement error. Holistic ratings also make it impossible to determine the basis for the examiners' ratings, and to separate examiner severity from candidate ability. If another examiner gives a holistic rating to the same candidate, they often differ significantly.

In an effort to gather more information about the candidate, the pertinent clinical skills encompassed in the holistic rating were broken out, and examiners were asked to give separate analytic ratings, one for each skill. The problem is how to collect enough information to make pass\fail decisions about candidates that have minimal measurement error and reasonable confidence in their accuracy, while not asking examiners for redundant ratings.

The medical skills tested in an oral certification examination, diagnosis, treatment, and technical skill are conceptually related by the nature of the clinical situation. This is why they are selected for use in the examination. Can these skills be evaluated independently by examiners in the examination environment. Is it possible to evaluate the choice of treatment independently from the diagnosis?

Candidates have an ability to perform the clinical skills. This ability is expected to be reasonably stable across time, skills and applications. The goal of the examination is to certify candidates as safe, competent physicians. If candidate performance on the examination across skills or across cases were extremely volatile, this would challenge the expectation that candidate competence represents a single meaningful construct.

It seems impossible for a candidate who received a low mark for the pivotal skill, diagnosis, to receive a high mark for treatment, since it would be highly unlikely that the candidate's inaccurate diagnosis would happen to have the same treatment as does the correct diagnosis. In fact, when skills are arranged in their clinical sequence, it should be unlikely that a higher grade would ever follow a lower grade. Therefore, skills conceptually arranged in clinical sequence, should show the same or consistently decreasing scores.

However, conceptual relation and lack of rating independence do not consider the relative difficulties of the skills. Relative skill difficulty levels result from the unique demands each skill requires. Skill difficulties are established independently of candidate abilities or examiner severities, with the Rasch multi-facet model (Linacre, 1989). Generally, candidates receive lower scores on more difficult skills and higher scores on easier skills, regardless of the clinical sequencing of the skills. When an easier skill is followed by a harder skill, candidates' scores are likely to decrease more often than not. Likewise, when a harder skill is followed by an easier skill, we expect candidates' scores to increase more often than not.

Data are from two different medical oral certification examinations. Skill ratings were given to candidates on a four point scale (EX1 scale = 1,2,3,4 and EX2 scale = 0,1,2,3). Both examinations were analyzed with the FACETS program (Linacre, 1990).

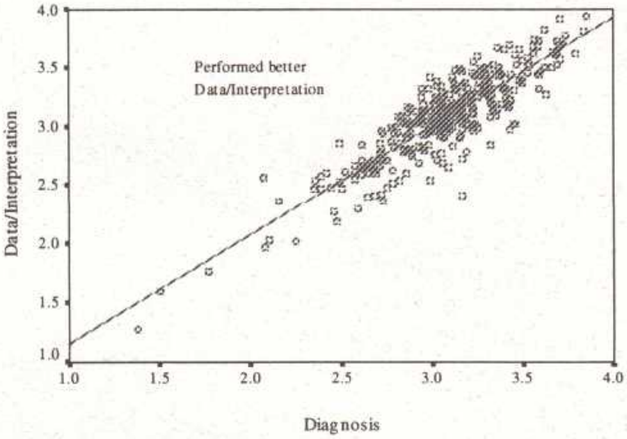
In the first examination, EX1, oral examiners rated candidates on three skills on each of four standardized cases. The skills were: 1) data /interpretation; 2) diagnosis; and 3) management. In this examination, examiners informed candidates of errors to insure that candidates continued through the standardized case as established. This examination is structured to minimize the effects of conceptual dependence and foster independent skills assessments. The second medical examination EX2, examined each candidate on cases from the candidate's actual practice. Candidates were rated on six skills: (1) data gathering; (2) diagnosis; (3) treatment; (4) technical skills (of surgery); (5) outcomes; and (6) ethics.

The FACETS program establishes a fair average score for each candidate on each skill. The fair average score is the score expectation of the logit measure and accounts for the severity of the examiner and difficulty of the standardized case. The fair average score is used in this analysis to make it easier to relate the scores to the rating scale. When fair average scores are the same for two skills, the ratings may not be independent, or the candidate may have the same level of ability on both skills. When the fair average scores differ, this suggests that examiners were able to distinguish candidate performance or that candidates demonstrated different levels of ability on each skill.

Table 1. Skill Difficulty Measures for EX1

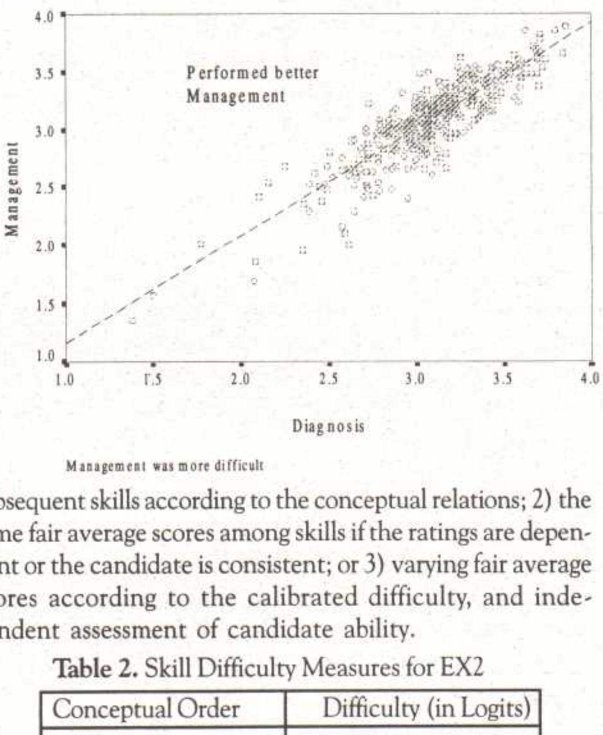
Conceptual Order	Difficulty (in logits)
Data Gathering	0.00
Diagnosis	-0.18
Treatment	0.18

Graph 1. Comparison of Performance on Two Skills



Data/Interpretation was more difficult

Graph 2. Comparison of Performance on Two Skills



Management was more difficult

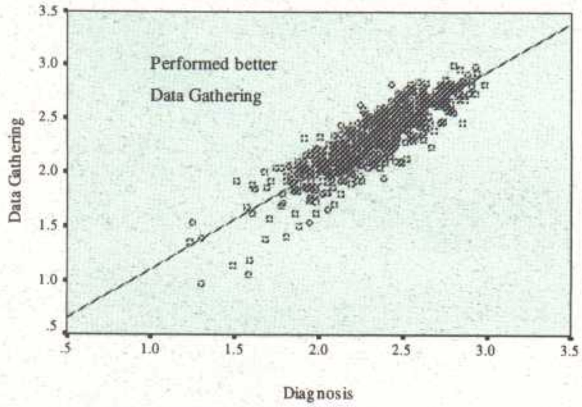
Diagnosis, a pivotal skill, is used for comparison to the other skills. Diagnosis is also a relatively easy skill for both EX1 and EX2, as shown in Tables 1 and 2. Therefore candidates should earn 1) the same or lower fair average scores on subsequent skills according to the conceptual relations; 2) the same fair average scores among skills if the ratings are dependent or the candidate is consistent; or 3) varying fair average scores according to the calibrated difficulty, and independent assessment of candidate ability.

Table 2. Skill Difficulty Measures for EX2

Conceptual Order	Difficulty (in Logits)
Data Gathering	.09
Diagnosis	–.21
Treatment	.16
Technical Skill	.08
Technical Skill	.05
Ethics	–.52

Table 1 shows the Rasch calibrated skill difficulties for EX1. Diagnosis is the easiest skill. Graphs 1 and 2 show the comparison of the fair average scores for diagnosis (easier) with data gathering (harder) or management (harder) respectively. Most candidates earned comparable fair average scores among skills, supporting the consistency of candidate ability among skills. However, some candidates earned higher or lower fair average scores on data/interpretation or management. This provides some evidence that examiners rate the skills independently based on their observation of the candidate and the difficultly of the skill.

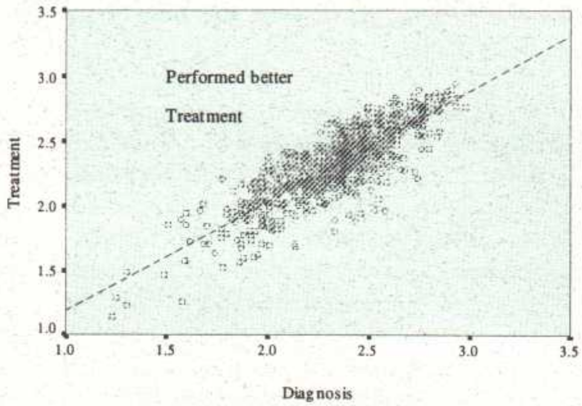
Graph 3. Comparison of Performance on Two Skills



Data Gathering was more difficult

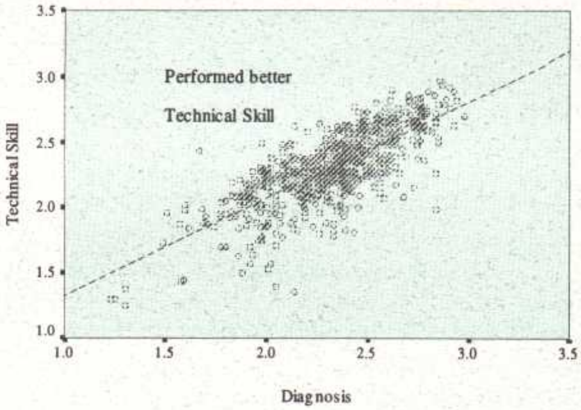
Table 2 shows the calibrated difficulties of the skills for EX2. Diagnosis is one of the easiest skills on which to earn a high score. Graphs 3 - 6 show the comparisons of fair average scores when diagnosis is compared to data gathering, treatment, technical skills, and outcomes respectively. Many of the candidates earn comparable fair average scores among skills. This is commensurate with the premise that candidates have a stable ability that can be measured. However, some candidates earn higher fair average scores on the clinically subsequent skills, showing that examiners can evaluate candidate performance, independent of the underlying conceptual relationships. These results show that the calibrated difficulty of the skill is not driven by conceptual relations among skills. While the functional relationship among the skills is critical to the coherence of the overall examination, the functional relationship does not control examiners' ratings. Rather, examiners seem to be able to rate candidates on each skill independently. This pattern holds true when examiners rate candidates on cases from their actual medical practices, or on standardized cases developed by the Board. The use of analytic ratings may not be foolproof, but examiners' analytic ratings appear to be independent, even when skills are conceptually related. In addition, the use of analytic rather than holistic ratings, has the advantage of collecting a sufficient amount of information about each candidate to make pass and fail decisions with minimal measurement error and a high level of confidence.

Graph 4. Comparison of Performance on Two Skills



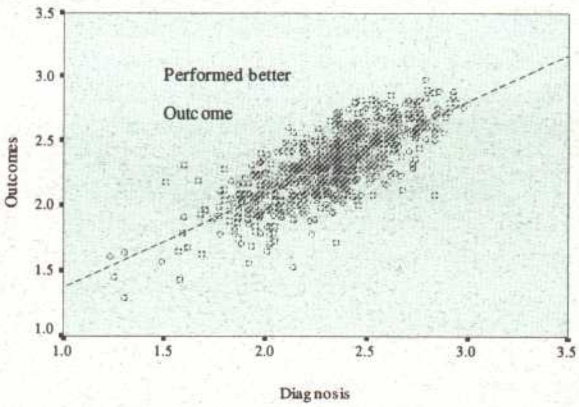
Treatment was more difficult

Graph 5. Comparison of Performance on Two Skills



Technical Skill was more difficult

Graph 6. Comparison of Performance on Two Skills



Outcomes was more difficult

References

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

Facets: A computer program. Chicago: MESA Press.

Thurstone's Crime Scale Re-Visited

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

In 1927, Louis Thurstone published a paper explicating the method of paired comparisons utilizing for this purpose the scaling of 19 criminal offenses. The purpose of his study was to further the cause of producing linear scales of social values. It was his lifelong task. The results of the 1927 study produced a crime scale that was replicated in order to determine how rankings of criminal offenses in 1927 compared to those of 1998, slightly more than that 70 years.

Thurstone chose these 19 offenses:

Abortion
Adultery
Arson
Assault and battery
Bootlegging
Burglary
Counterfeiting
Embezzlement
Forgery
Homicide
Kidnapping
Larceny
Libel
Perjury
Rape
Receiving stolen goods
Seduction
Smuggling
Vagrancy

Method

Thurstone arranged the criminal offenses so each was paired with each of the other listed offenses. This produces n (n - 1) = 171 pairs. He administered the list to 266 students at The University of Chicago. In preliminary work, Thurstone found that some college students were not familiar with various terms, so he provided a sheet of definitions.

Sample

I used the same set of pairs and with the assistance of students in my psychometrics classes, administered the 171 pairs with the same definitions to a large number of samples including, for the sake of this comparison, 260 college students. As near as I can determine, my study replicated his methodology in sample size and composition. The materials used were exactly the same.

Results

Paired comparisons for the 19 offenses produces a large array of items, 171, which presents a considerable task to each subject, but an even greater task when tabulated by hand and transformed from individual responses to tally sheets, subsequently totaled and converted from proportions to a linear scale.

The development of linear scaling was a goal of Thurstone and the method of paired comparisons was one of the techniques he used. The method of equal interval scaling is another of his methods. But like the method of equal appearing intervals, paired comparisons especially when computed by hand, requires much time and detailed effort. It is no surprise that these onerous methods are ignored in favor of simpler methods such as Likert scaling. [However, I might add that we are the losers in social science for this neglect and that the process can be greatly simplified with the use of computer software. Using BIGSTEPS and WINSTEPS greatly reduces the labor and produces a Rasch analysis of the scale.

There are several ways the comparisons between the two samples might be made. Fortunately, Thurstone provided scale values for the 1927 scale to which the current values could be compared. These values are given in Table 1. A scatter plot of the 19 points for each of the criminal offenses is most revealing. Figure 1 gives a plot of the criminal offenses numbering the offenses in the order presented in Table 2. The correlation between the two sets of values is 0.51 significant beyond the .05 level. The 95% control lines indicate that almost all of the data points are within and only item 11, bootlegging, is an outlier. Using 68% control lines, not shown, item 17, seduction, and item 19, vagrancy, are outliers but each one is only slightly above and below the 68% control lines respectively.

Discussion

The remarkable similarity in scaling criminal offenses by two similar samples of college students and separated by 70 years appears remarkable. The availability of Thurstone's methodology and resulting scale values, allowed the comparison to be more exact that many sample comparison are

Figure 1 Map of crimiinal offenses CRIMERAT.MHS

Scale Value	1927 data	1966 data	1999 date
100	Rape	Homicide	Homicide
	Homicide		
95			
90			Kidnapping
			Rape
		Rape	
85			
80			
		Kidnapping	
75			
70	Abortion,Seduction		
	Kidnapping		
65		Arson	
	Adultery	Assault-battery	Perjury
	Arson		
60			
			Assault-battery,Counterfeiting
			Arson,Forgery
55			
	Perjury		
50	Embezzlement		
	Counterfeiting		Abortion,Adultery,Smuggling
	Burglary,Forgery		Libel
45	Assault-battery	Abortion,Burglary	
		Embezzlement	
40	Larceny	Adultery,Perjury	
		Counterfeiting,Larceny	
		Seduction	
35			
	Smuggling,Libel	Forgery	
	Bootlegging		Bootlegging,Burglary
30	Receiving stolen goods	Smuggling,Libel	Receiving stolen goods
		Embezzelment	
25		Seduction	
20			
		Receiving stolen goods	
15			
		Bootlegging	
10			
			Larceny
5			
0	Vagrancy	Vagrancy	Vagrancy

over the space of such a period of time. The general liberality of college age students compared to adults is not a factor of this study, but one cannot help but be struck by the similarity in scaling criminal offenses for this age group. The results suggest that the ranking of criminal offenses has not undergone any substantial changes for this period of time for this age group. Bootlegging, understandably so, rated higher in the late 1920's than it does today. Seduction was rated higher in the earlier sample than among current students; the recent news coverage of "sexual" matters in the nation's capitol does not make this difference surprising. More recent coverage of criminal reporting in the media,often in connection with politicians whose behavior appears to be under increased scrutiny, has not substantially changed students' perceptions of criminal offenses except for those already noted.

Methodology may play a positive part in these results. It is fortunate that a researcher of Thurstone's stature was involved in the initial study. His work was thorough, complete and easy to follow. These are traits important in social science research. Replication was relatively easy. It is important to know whether or not social values are stable. If there is change, the researchers need to be aware of the change in direction and the degree of the change. Social values are intangible and not easy to determine. People have strong feelings about crime and recent coverage in the media has, perhaps, polarized opinions as against reasoned scrutiny of values and their origins. These findings suggest that there is surprising stability in college students' perceptions of the seriousness of criminal offenses.

References

Thurstone, L. L. (1927). The method of paired comparisons for social values. Journal of Abnormal and Social Psychology, 21, 384-400.

Thurstone, L. L & Chave, E.. (1929). The measurement of attitude. Chicago: The University of Chicago Press.

Torgerson, W.S. (1958). Theory and methods of scaling. New York: Wiley.

Toward a Definition of Sexual Harassment in the Workplace

Suzy Vance JD, LL.M.
Anne Wendt, PhD, RN

Introduction

Reports of sexual harassment on the job are on the rise nationwide. Employers are seeking strategies to decrease and prevent sexual harassment. This report is based on: 1) training work sessions intended to increase participants' self-awareness and appreciation of others, and 2) assessment of shifts in participants' attitudes and awareness with respect to potential sexual harassment behaviors. The unique work sessions consisted of:

1) presentation and discussion of information about what constitutes sexual harassment,
2) group activities intended to raise awareness of self and others, and
3) presentation of scenarios portraying common work situations using live actors and volunteers from among the participants to build skills for managing human interactions in the workplace.

Methodology

Participants were asked to complete a survey assessing sexual issues/harassment before and after the work sessions. After a review of the literature and legal cases relating to sexual issues/harassment, the authors developed a theory about how sexual issues/harassment might be manifested in the workplace. Table 1 illustrates the spectrum of potentially problematic behavior in the area of sexual issues/harassment.

Table 1. Spectrum of Behavior

Visual	Verbal	Written	Touching	Power	Force
Staring	Requests for dates	Love letters	Violating space	Using position to insist on dates and other things	Rape
Posters	Lewd comments	Obscene letters	Patting		Physical
Magazines	Sex jokes	Cards	Grabbing	Promising	
assault					
Calendars	Questions about personal life	E-mail Fax	Caressing Kissing Fondling	Threatening with negative impact on job	

* This survey instrument, an Interfocus® Survey - Human Interaction in the Workplace #1, has been registered with the Copyright Office of the Library of Congress by Susan Vance.

Suzy Vance

Over the years Suzy, with her common sense and sensitivity to diverse perspectives, engaged in an extensive and successful law practice focusing on human relationships at work - including the presentation of a prevailing argument to the United States Supreme Court.

Today Human Interaction is her business. Her work with groups and organizations is based on her fundamental belief that: "People make the difference in all we do."

Suzy offers services in three areas: Interfocus® building strategies for human interaction in the workplace while addressing specific concerns. Partnership Connection® - bridging the gap from school to community through inter-generational programs in elementary, middle and secondary schools. Team-building - Bonding and strengthening groups and rewarding people for jobs well done - including Life Mask®.

Instrument Development

A survey intended to illicit honest responses from participants regarding sexual issues/harassment was developed. Participants were asked to respond to statements using a likert-type scale in the following areas: jokes, flirting, dress and attraction, touching, patting, hugging, and backrubs.

The survey began with "easy-to-agree with" statements that are playful and engaging. The statements become "harder-to-agree with" and more risky and dangerous for workplace behavior. For example, it is "safe" and "easy-to-agree with" the statement "I laugh at good jokes." It is "riskier" and "harder-to-agree with" the statement "I like to tell sex jokes." Similarly, it is logical that it is "safe" and "easy-to-agree with" the statements "I like back rubs" and "I like getting back rubs." It is "more risky" and "harder-to-agree with" the statement "Backrubs at work are ok."

An example of how the statements were formatted in the survey is as follows:

1.	I enjoy sex jokes.	SA	A	D	SD
2.	I tell sex jokes.	SA	A	D	SD
3.	Sex jokes are ok, as long as they don't stop work.	SA	A	D	SD

SA = Strongly Agree, A = Agree, D = Disagree, SD = Strongly Disagree

Data Collection

For reasons peculiar to the project, demographic information was not collected. There is no information as to the differences in response, if any, between men and women, various levels within the department, or racial or ethnic distinctions. There also is no information as to the movement on the spectrum or shift in responses for individual participants because responses were not tracked individually. Without demographic information, the results reported here represent only a beginning definition of the variable "Sexual Issues/Harassment."

Surveys were distributed to 216 participants before training. One hundred and eighty five of the 216 employees attended the first work session. One hundred one (55%) of the 185 participants turned in the "before" survey. One hundred sixty seven of the 216 employees attended the second work session. One hundred eleven (66%) of the 167 completed the "after" survey. Twenty six (12%) surveys were determined to be invalid. One hundred eighty six surveys were analyzed to determine the definition of the variable.

Data Analysis

Data were analyzed using the Rasch partial credit model with WINSTEPS. Data which did not fit the model were not used as part of the definition of the Sex Issues Construct.

Results

The responses to the before-and-after surveys were pooled to create the Sex Issues Construct shown in Table 2. Analysis of the data shows that responses fell into three categories - statements that were

1. **SAFE** - Easiest to agree with - more than 50% agreed
2. **RISKY** - Easier to disagree with - more than 50% disagreed
3. **DANGEROUS** - Very Much Easier to disagree with - more than 67% disagreed

Table 2. Sex Issues Construct

SAFE - Easiest to AGREE with

I laugh at good jokes.
Jokes at work are ok.
How much I enjoy being hugged, depends on the circumstances.
When I kid around, I might pat someone on the back.
I like to tell jokes.
When I congratulate someone, I pat them on the back.
It's ok to hug a co-worker.
I like back rubs.
I like getting back rubs.
I like to hug.
Sometimes I touch people without knowing it.
When I'm excited, I might hug.

RISKY - Easier to DISAGREE with

When someone wears an outfit, I may stare at them.
When someone dresses in an appealing way, I like to tell them.
I like to touch people.
It's ok to hug the boss.
Whether I enjoy a sex joke, depends on who tells it.
I enjoy sex jokes.
Touching at work is ok.
When I see something I want, I go after it.
I like giving back rubs.
I enjoy flirting.

DANGEROUS: Much Easier to DISAGREE with

Worrying about "not touching" is silly.
Flirting at work is ok.
Sex jokes are ok at office parties.
When I think someone is good-looking, I let them know.
Sex jokes are ok, as long as they don't stop work.
When I'm attracted to someone, I'm not afraid to tell them.
Flirting is ok as long as it doesn't stop work.
When I want to go out with someone, I ask them.
Flirting is harmless.
Flirting never makes me uncomfortable.
I tell sex jokes.
I like to flirt at work.
Back rubs are ok at work.
When someone is good looking, I can't stop looking at them.
When I kid around, I might pat someone on the rear.

For example, statements about flirting were "harder-to-agree with" and thought to be risky and dangerous. It was "hardest-to-agree" that patting someone on the rear is ok. "Patting on the rear" is the most dangerous of all identified interactions and borders on the more overt end of the behavior spectrum constituting portential sexual harassment.

(**Note:** This does not mean there should be a rule against getting or giving back rubs or hugs at work, flirting at work, or even patting someone on the rear. What the Sex Issues Construct does show is attitudes toward certain behavior fall in a logical or commonsense progression from most "safe" to most "dangerous." This information can be used to measure shifts in awareness and appreciation or attitude. It also can be used to raise awareness of the progression of behavior and teach skills to avoid or stop the progression when it becomes important to prevent behavior from crossing the line from "safe" into "risky" and "dangerous" areas.)

As can be seen in Table 2, the goal of the authors of determining the Sex Issues Construct was achieved. Those statements on attitudes and behaviors which were intended to be "easy-to-agree with," such as "I laugh at good jokes," have indeed calibrated to be safe and "easy-to-agree with." Those statements which were intended to be "harder-to-agree with" such as "I tell sex jokes" and "Backrubs at work are ok" have indeed calibrated to be "dangerous" and "more difficult to agree with."

Conclusions

The initial definition of the Sexual Issues Construct essentially has been realized.

There were several statements on the survey that did not fit the Rasch measurement model. They were not used in the definition of the Sex Issues Construct. These statements will need to be revised as the definition of the Sex Issues Construct is refined. Some statements also did not fit within the Sex Issues Construct as the authors of the survey anticipated. For example, "Flirting never makes me uncomfortable" was not thought to be one of the statements most "hard-to-agree with." The use of the modifying word "never" in the statement may have contributed to this unanticipated result. Additional statements also need to be created to fill in gaps and expand the continium of the Sex Issues Construct.

Anne Wendt

Anne Wendt is the NCLEX Content Manager at the National Council of State Boards of Nursing, a not-for-profit organization responsible for the development of the National Council nursing licensure examination (NCLEX Examination). She received her BSN from the University of Minnesota, her MSN from Loyola University, and her Ph.D. in Psychometrics from the University of Chicago.

Anne Wendt has a unique perspective of nursing licensure exams because she comes to her position as a nurse, a psychometrician, and as an educator. She was instrumental in the National Council's transition from a paper-and-pencil NCLEX examination to its current computerized adaptive testing (CAT) form. She has co-authored the NCLEX test plans and detailed test plans since March 1993. She has also been influential in the publication of such documents as The NCLEX™ Process, The NCLEX™ Manual and Assessment Strategies for Nursing Educators.

SURVEY DESIGN RECOMMENDATIONS

William P. Fisher, Jr., Ph.D.
Public Health & Preventive Medicine
LSU Health Sciences Center - New Orleans

Item writers and data analysts should follow seventeen basic rules of thumb to create surveys that

1) are likely to provide data of a quality high enough to meet the requirements for measurement specified in a probabilistic conjoint measurement (PCM) model;
2) implement the results of the PCM tests of the quantitative hypothesis in survey and report layouts, making it possible to read interpretable quantities off the instrument at the point of use with no need for further computer analysis; and
3) are joined with other surveys measuring the same variable in a metrology network that ensures continued equating (Masters, 1985) with a single, reference standard metric

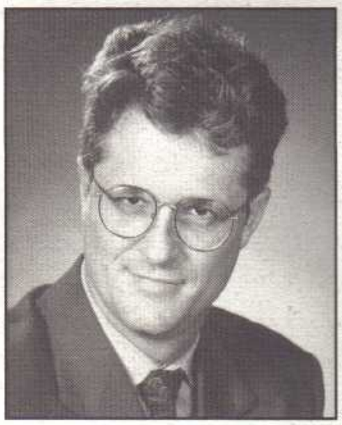
First, make sure all items are expressed in simple, straightforward language.

Second, restrict each item to one idea, meaning avoid conjunctions (and, but, or), synonyms, and dependent clauses. A conjunction indicates the presence of at least two ideas in the item. Having two or more ideas in an item is unacceptable because there is no way to tell from the data which single idea or combination of ideas the respondent was dealing with. If two synonymous words really mean the same thing, only one of them is needed. If the separate ideas are both valuable enough to include, they need to be expressed in separate items. Dependent (if, then) clauses require the respondent to think conditionally or contingently, adding an additional and usually unrecoverable layer of interpretation behind the responses that may muddy the data.

Third, avoid "Not Applicable" or " No Opinion" response categories. It is far better to instruct respondents to skip irrelevant items than it is to offer them the opportunity in every item to seem to provide data, but without having to make a decision.

Fourth, avoid odd numbers of response options. Middle categories tend to attract disproportionate numbers of responses. Again, it allows the respondent to appear to be providing data, but without making a decision concerning preferences. If someone really cannot decide which side of an issue they come down on, let them decide on their own to skip the question.

Fifth, do not assume that respondents will be unable to make more than one or two distinctions in their responses, and do not simply default to the usual four response options (Strongly Agree, Agree, Disagree, Strongly Disagree, or Never, Sometimes, Often, and Always, for instance). The LSU HSI PFS,(Fisher, Marier,



William P. Fisher, Jr.

William P. Fisher, Jr. was formerly Senior Research Scientist for Program Evaluation at Marianjoy Rehabilitation Hospital & Clinics in Wheaton, IL, serving on the Management Team, and on the Clinical Programs and Quality Assessment & Improvement Committees. After completing the University of Chicago's Social Sciences Divisional Master's degree in 1984, William was a Spencer Foundation Dissertation Fellow, earning a Ph.D. in Chicago's Department of Education in 1988, concentrating in Measurement, Evaluation, and Statistical Analysis (MESA). Dr. Fisher is still a MESA Research Associate, is on the Editorial Board of the Journal of Outcome Measurement, and on the Advisory Board of the Institute for Objective Measurement. He is professionally active in diverse organizations. Current tasks include designing and implementing an outcome measurement system for Louisiana's statewide public hospital system; consulting on the Social Security Administration's Disability Process Redesign Project; and drafting scale-free health status measurement standards for the ASTM E31 Committee on Medical Informatics.

Eubanks & Hunter, 1997; Fisher, Eubanks & Marier, 1997) for example, employs a six-point rating scale and is intended for use in the Louisiana statewide public hospital system, which provides most of the indigent care in the state. To date, about 75% of the respondents have less than a high school education and incomes of less than $15,000 per year, but they have shown little or no difficulty in providing consistent responses to the questions posed. Part of the research question raised in any measurement effort concerns determining the number of distinctions that the variable is actually capable of supporting, besides determining the number of distinctions actually required for the needed comparisons. Starting with six (adding in Very Strongly Agree/Disagree categories to the ends of the continuum) or even eight (adding Absolutely Agree/Disagree extremes) response options gives added flexibility in survey design. If one or more categories blends with another and isn't much used, the categories can be combined. Research that starts with fewer categories, though, cannot work the other direction and create new distinctions. More categories have the added benefit of boosting measurement reliability, since, given the same number of items, an increase in the number of functioning (used) categories increases the number of distinctions made among those measured.

Sixth, write questions that will provoke respondents to use all.of the available rating options. This will maximize variation, important for obtaining high reliability.

Seventh, write enough questions and have enough response categories to obtain an average error of measurement low enough to provide the needed measurement separation reliability, given sufficient variation. Reliability is a strict mathematical function of error and variation and ought to be more deliberately determined via survey design than it currently is (Linacre, 1993; Woodcock, 1992). For instance, if the survey is to be used to detect a very small treatment effect, measurement error will need to be very low relative to the variation, and discrimination will need to be focused at the point where the group differences are effected, if statistically significant and substantively meaningful results are to be obtained. On the other hand, a reliability of .70 will suffice to simply distinguish high from low measures. Given that there is as much error as variation when reliability is below .70, and it is thus not possible to distinguish two groups of measures in data this unreliable, there would seem to be no need for instruments in that range.

Eighth, before administering the survey, divide the items into three or four groups according to their expected scores. If any one group has significantly fewer items than the others, write more questions for it. If none of the questions are expected to garner very low or very high scores, reconsider the importance of step six above.

Ninth, order the items according to their expected scores and consider what it is about some questions that make them easy (or agreeable or important, etc.), and what it is about other questions that make them difficult (or disagreeable, unimportant, etc.). This exercise in theory development is important because it promotes understanding of the variable. After the first analysis of the data, compare the empirical item order with the theoretical item order. Do the respondents actually order the items in the expected way? If not, why not? If so, are there some individuals or groups who did not? Why?

Tenth, consider the intended population of respondents and speculate on the average score that might be expected from the survey. If the expected average score is near the minimum or the maximum possible, the instrument is off target. Targeting and reliability can be improved by adding items that provoke responses at the unused end of the rating scale. Measurement error is lowest in the middle of the measurement continuum, and increases as measures approach the extremes. Given a particular amount of variation in the measures, more error reduces reliability and less error increases it. Well-targeted instruments enhance measurement efficiency by providing lower error, increased reliability, and more statistically significant distinctions among the measures for the same number of questions asked and rating options offered.

Eleventh, as soon as data from 30-50 respondents are obtained, analyze the data and examine the rating scale structure and the model fit using a partial credit PCM model. Make sure the analysis was done correctly by checking responses in the Guttman scalogram against a couple of respondents' surveys, and by examining the item and person orders for the expected variable. Identify items with poorly populated response options and consider combining categories or changing the category labels. Study the calibration order of the steps and make sure that a higher category always represents more of the variable; consider combining categories or changing the category labels for items with jumbled step structures. Test out recodes in another analysis; check their functioning, and then examine the item order and fit statistics, starting with the fit means and standard deviations in BIGSTEPS Table 3. If some items appear to be addressing a different construct, ask if this separate variable is relevant to the measurement goals. If not, discard or modify the items. If so, use these items as a start at constructing another instrument. When the step structure and model fit are orderly, either continue gathering data on the existing survey and be prepared to make the same edits and changes later with more data, or modify the survey and gather new data in the new format.

Twelfth, when the full calibration sample is obtained, maximize measurement reliability and data consistency. First identify items with poor model fit. If an item is wildly inconsistent, with a mean square fit statistic markedly different from all others, examine the item itself for reasons why its responses should be so variable. Does it perhaps pertain to a different variable? Does the item ask two or more very different questions at once? It may also be relevant to find out which respondents are producing the inconsistencies, as their identities may suggest reasons for their answers. If the item itself seems to be the source of the problem, it may be set aside for inclusion in another scale, or for revision and later re-incorporation. If the item is functioning in different ways for different groups of respondents, then the data for the two groups ought to be separated into different columns in the analysis, making the single item into two. Finally, if the item is malfunctioning for no apparent reason and for only a very few otherwise credible respondents, it may be necessary to omit only specific, especially inconsistent responses from the calibration. Then, after the highest reliability and maximim data consistency are achieved, another analysis should be done, one in which the inconsistent responses are replaced in the data. The two sets of measures should then be compared in plots to determine how much the inconsistencies actually affect the results.

Thirteenth, the instrument calibration should be compared with calibrations of other similar instruments used to measure other samples from the same population. Do similar items calibrate at similar positions on the measurement continuum? If not, why not? If so, how well do the pseudo-common items correlate and how near the identity line do they fall in a plot? If the rating scale step structures are different, are the step transition calibrations meaningfully spaced relative to each other?

Fourteenth, the calibration results should be fed back onto the instrument itself. When the variable is found to be quantitative and item positions on the metric are stable, that information should be used to reformat the survey into a self-scoring report. This kind of worksheet makes it possible to build the results of the instrument calibration experiment into the way information is organized on a piece of paper, providing quantitative results (measure, error, percentile, qualitative consistency evaluation, interpretive guidelines) at the point of use. No survey should be considered a finished product until this step is taken.

Fifteenth, data should be routinely sampled and recalibrated to check for changes in the respondent population that may be associated with changes in item difficulty.

Sixteenth, for maximum utility, the instrument should be equated with other instruments intended to measure the same variable, creating a reference standard metric.

Seventeenth, everyone interested in measuring the variable should set up a metrology system, a way of maintaining the reference standard metric via comparisons of results across users and brands of instruments. To ensure repeatability, metrology studies typically compare measures made from a single homogeneous sample circulated to all users. Given that this is an unrealistic strategy for most survey research, a workable alternative would be to occasionally employ two or more previously equated instruments in measuring a common sample. Comparisons of these results should help determine whether there are any needs for further user education, instrument modification, or changes to the sampling design.

References

Andrich, D. (1988). Rasch models for measurement. Sage University Paper Series on Quantitative Applications in the Social Sciences, vol. series no. 07-068. Beverly Hills, California: Sage Publications.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. Psychometrika, 42, 631-634.

Connolly, A. J., Nachtman, W., & Pritchett, E. M. (1971). Keymath: Diagnostic Arithmetic Test. Circle Pines, MN: American Guidance Service.

Fisher, W. P., Jr. (1996, October). Rating scale measurement standards relevant to ASTM 1384 on the content and structure of the electronic health record. Unpublished paper. ASTM E31 Committee on the Electronic Health Record, Washington, DC.

Fisher, W. P., Jr. (1997a). Physical disability construct convergence across instruments: Towards a universal metric. Journal of Outcome Measurement, 1(2), 87-113.

Fisher, W. P., Jr. (1997b, June). What scale-free measurement means to health outcomes research. Physical Medicine & Rehabilitation State of the Art Reviews, 11(2), 357-373.

Fisher, W. P., Jr. (1998, May). Objectivity in psychosocial measurement: What, why, how. Second International Outcome Measurement Conference. University of Chicago.

Fisher, W. P., Jr. (1999). Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. Journal of the Louisiana State Medical Society (submitted).

Fisher, W. P., Jr., Eubanks, R. L., & Marier, R. L. (1997). Equating the MOS SF36 and the LSU HSI physical functioning scales. Journal of Outcome Measurement, 1(4), 329-362.

Fisher, W. P., Jr., Harvey, R. F., & Kilgore, K. M. (1995). New developments in functional assessment: Probabilistic models for gold standards. NeuroRehabilitation, 5(1), 3-25.

Fisher, W. P., Jr., Marier, R. L., Eubanks, R., & Hunter, S. M. (1997). The LSU Health Status Instruments (HSI). In J. McGee, N. Goldfield, J. Morton & K. Riley (Eds.), Collecting Information from Patients: A Resource Manual of Tested Questionnaires and Practical Advice (Supplement) (pp. 13:109-13:127). Gaithersburg, Maryland: Aspen Publications, Inc.

Fisher, W. P., Jr., & Wright, B. D. (1994). Introduction to probabilistic conjoint measurement theory and applications. International Journal of Educational Research, 21(6), 559-568.

Linacre, J. M. (1993). Rasch generalizability theory. Rasch Measurement Transactions, 7(1), 283-284.

Linacre, J. M. (1997). Instantaneous measurement and diagnosis. Physical Medicine and Rehabilitation State of the Art Reviews, 11(2), 315-324.

Mandel, J. (1977, March). The analysis of interlaboratory test data. ASTM Standardization News, 5, 17-20, 56.

Mandel, J. (1978, December). Interlaboratory testing. ASTM Standardization News, 6, 11-12.

Masters, G. N. (1985, March). Common-person equating with the Rasch model. Applied Psychological Measurement, 9(1), 73-82.

Masters, G. N., Adams, R. J., & Lokan, J. (1994). Mapping student achievement. International Journal of Educational Research, 21(6), 595-610.

O'Connell, J. (1993). Metrology: The creation of universality by the circulation of particulars. Social Studies of Science, 23, 129-173.

Pennella, C. R. (1997). Managing the metrology system. Milwaukee, WI: ASQ Quality Press.

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

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Sparrow, S., Balla, D., & Cicchetti, D. (1984). Interview edition, survey form manual, Vineland Behavior Scales. Circle Pines, MN: American Guidance Services, Inc.

Suppes, P., Krantz, D., Luce, R., & Tversky, A. (1989). Foundations of measurement, Volume II: Geometric and probabilistic representations. New York: Academic Press.

Wise, M. N. (Ed.). (1995). The values of precision. Princeton, NJ: Princeton University Press.

Woodcock, R. W. (1973). Woodcock Reading Mastery Tests. Circle Pines, MN: American Guidance Service, Inc.

Woodcock, R. W. (1992). Woodcock test design nomograph. Rasch Measurement Transactions, 6(3), 243-244.

Woodcock, R. W. (1999). What can Rasch-based scores convey about a person's test performance? In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement: What every psychologist and educator should know. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), Measurement and personality assessment. North Holland: Elsevier Science Ltd.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement: What every educator and psychologist should know. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rasch Analysis for Surveys

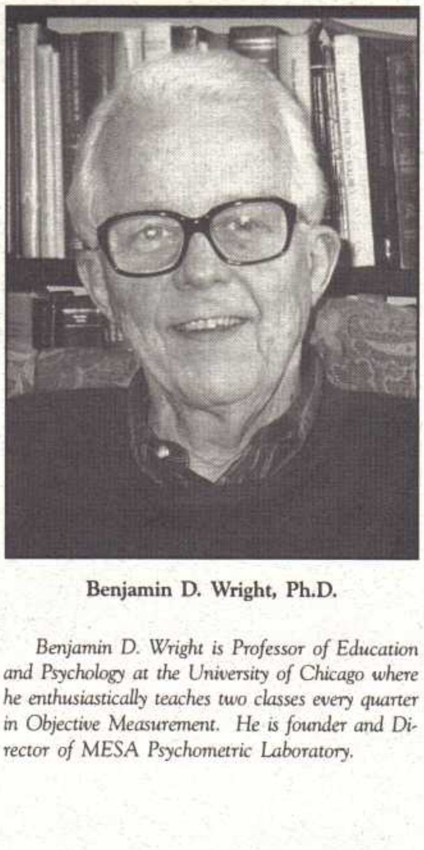
Ben Wright

Surveys, questionnaires and interview protocols that use rating scales to collect psychosocial information can be thought of as structured "conversations" between researchers and subjects. To construct a successful questionnaire, the researcher must develop a clear idea of the aim of the questionnaire, especially the inferences that are to be drawn from its use. The researcher must also be intimate with the language the intended subjects understand and use. Observed responses are local descriptions of a situation as perceived by the subject at a moment in time. From these passing responses, the researcher hopes to induce general inferences concerning reproducible processes of enduring psychosocial significance. The desired generalization requires that the observed responses can be fit into an overall metric, linear variable, along with moreness and lessness have well defined quantitative and qualitative meanings. The Rasch Model meets these criteria.

Rasch analysis is a method for constructing linear system from observed counts and categorical responses (like Likert scales), within which items and subjects can be measured unambiguously. The constructed variables contain the meaning of the structured "conversations." The measure of a subject on each variable summarizes that subject's statements about the variable to the extent that the subject shares a definition of the variable with other correspondents. These measures are the most succinct and reproducible report of the information collected by the questionnaire.

Rasch analysis facilitates the transmission of results to subsequent analyses, but now with the advantage of being linear measures with standard errors of the kind required by most statistical analyses. It also simplifies communication of results to therapists, educators, policy makers and the concerned public, in the form of graphical summaries of client populations and detailed individual client profiles.

A unique asset of Rasch analysis is its ability to detect idiosyncrasies - particular, specific departures of subjects and items from the shared understanding that is emerging from the ongoing research. These local departures have powerful diagnostic implications for the treatment of individual subjects. They also suggest new insights into the nature of the proposed variable and new possibilities for improving its definition and measurement.



Benjamin D. Wright, Ph.D.

Benjamin D. Wright is Professor of Education and Psychology at the University of Chicago where he enthusiastically teaches two classes every quarter in Objective Measurement. He is founder and Director of MESA Psychometric Laboratory.

Expert Panels, Consumers, and Chemistry

Thomas K. Rehfeldt

> We must next consider what account we are to give of any one of them; what, for example, we should say color is, or sound, or odor, or savor; and so also respecting [the object of touch. . . The point of our present discussion is, therefore, to determine what each sensible object must be in itself, in order to be perceived as it is in actual consciousness.
>
> Aristotle, (c330 B.C.) "On Sense and the Sensible"

IT'S THE ECONOMY

Large amounts of time, resources, and money are spent each year in the development of consumer products. Very large expenditures spent needlessly if the consumer does not like the products once in the market place. Thus, many more dollars are spent on consumer research to learn if the products will be embraced when on the market.

The process of chemical development is usually followed by expert panel evaluation, then one or more small consumer surveys, followed by a full-scale market research study. Anything that can be done to make the process more efficient, and, particularly, to make the testing predictive of consumer behavior is extremely valuable.

PREDICT WHAT?

Conventional market research testing makes use of methods such as factor analysis, multi-dimensional scaling, discriminant function analysis, and such like complex statistical methods. The value and advancement in these techniques, especially since the advent of cheap computing, has greatly increased in recent years. But, given the value and prevalence of these methods, one item is still lacking. These methods are not measurements and thus are not predictive, but solely descriptive of the most recent data.

Thomas K. Rehfeldt

Thomas K. Rehfeldt is Senior Research Statistician at the Unilever Home and Personal Care Innovation Center in Rolling Meadows, IL. After many years as an analytical chemist and developer of high tech. aircraft paint, he decided that he could have just as much fun solving problems that did not necessitate getting contaminated with smelly chemicals. He started in the statistics program at the University of Chicago where a fortuitous meeting with Ben Wright started him on the measurement path. Since then, he has tried to apply Rasch measurement tools in product development and marketing applications. He is currently working on the measurement of sensory perception as it relates to consumer choice and definition. His research interests include defining perception variables, relating expert panel judgment, consumer, marketing, and chemical data, and developing useful predictive models of consumer behavior. All of these entail analysis of rating scale, ranks, and dichotomous responses in many facets. When not chained to his computer he can be found building a new house in the north woods of Wisconsin or in the company of the two dogs to whom he belongs.

In the parlance of marketers, the predictions needed are "What are the key drivers of product acceptability?" and "What change in key drivers will produce a proportional change in acceptability?" The key drivers are those attributes, out of all possible product properties, that are the ones that are necessary for product acceptability, e.g., a shampoo may clean hair, but it will not sell if it does not lather. The key drivers may also be the complementary attributes; those that will cause the product to be rejected independently of the others, e.g., a shampoo may do everything well but have an undesirable fragrance.

In this context, the objective is to know what can be measured that will inform us of the effect of these key drivers, and what other facets may predict the level of acceptance.

OUR EXAMPLE

The example presented is for an examination of the attributes properties and acceptance of anti-perspirant products. Fourteen commercial products were tested in the consumer test; 400 consumers used 3 products, sequentially, for 2 weeks each. In the expert descriptive panel test, each of 14 panelists tested all products and 2 replicate trials were made. Analytical instrumental testing measured lightness, friction and rate of application for 14 products.

The objective is to identify the key drivers for the consumers. From history and experience, the drivers would be the efficacy, i.e., how it protects from odor and wetness, and application, i.e., how it feels when applied.

THE MEASUREMENT

Three types of data were collected: the analytical data, the expert panel data and the consumer data. The analytical data is a continuous scale. The whiteness was measured with a spectrophotometer; this is the L–value. The force to pull the anti-perspirant stick across a test material was measured as the dynamic friction. The consumer data was from a 10-point categorical scale, i.e., subjects were not allowed to mark fractional values but would check boxes at each of the scale marks. The expert panel data was collected as a 10 point continuous scale, i.e., subjects were allowed to mark the scale at any place on the line from 1 to 10. The direction of the consumer scale was cast as level of approval, so the direction was the same for all attributes. The expert panel data is collected as amount of the attribute so the direction of preference is not the same for all attributes.

This set of conditions illustrates the power of the Rasch model. Based on the set of common products, the expert panel data and the consumer data can be combined. The difference in how we ask the questions is of little concern. Since linear continuous measures are calculated, the analytical data is easily combined with the measures.

THE RESULTS

First, a FACETS analysis was performed on the consumer data. The FACETS program was used to account for the different attributes, the different products, the subjects and the replication or order of presentation. The measures for the attributes were examined. In the consumer study, all of the questions are worded so that all of the attribute scores progressed in the same direction. The questions were generally "How did you like the attribute?"

It was found that negative attributes scored high, along with positive attributes, indicating that the approved rating was related to lack of something (like greasiness).

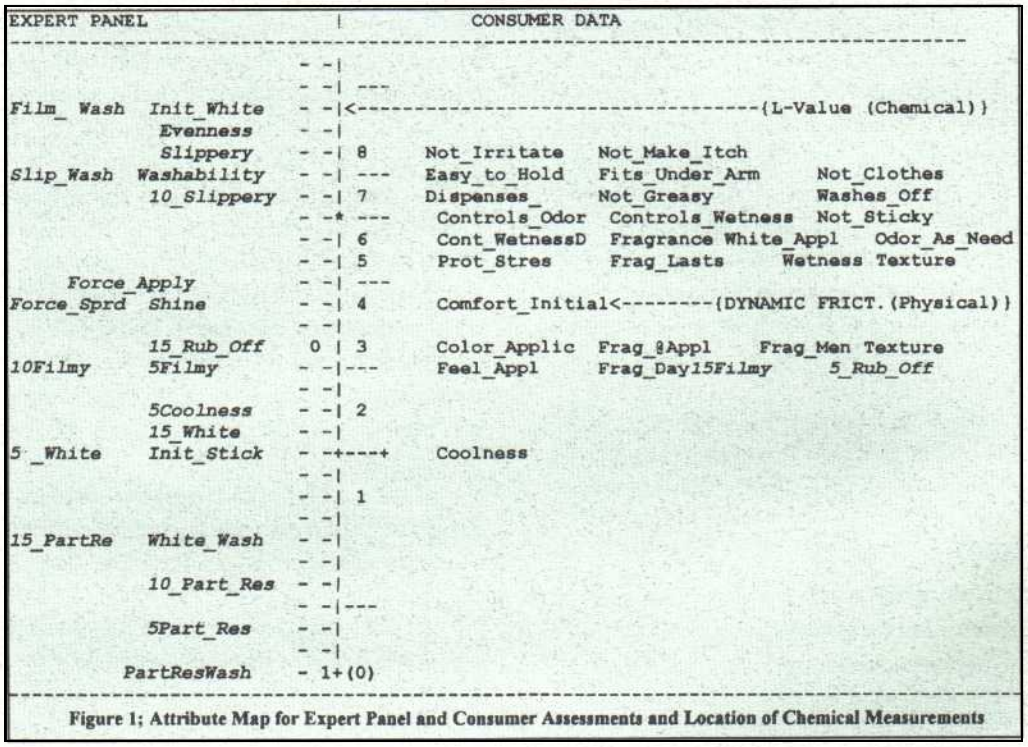
The next step was to run a FACETS analysis on the expert panel data. The results obtained were similiar to the consumer data, with the products serving as the common link. The expert panel is trained to report the amount of an attribute on the 10 point continuous scale; no distinction is made for undesirable attributes. Low greasiness was reported as 'less grease'.

The first comparisons found some of the attributes, on opposite ends of the scale, due to different form of the questions, i.e., greasiness was generally low for commercial products, so the expert panel reported low greasiness, which produced the consequent low measures. In contrast, low greasiness is seen as desirable to the consumer so they approved this and gave a high approval score.

By judicious choice of the centering and anchoring, the expert panel measurements and the consumer measurements are on the same scale and in the same direction. The final measurement scales are shown in Figure 1.

One can observe which expert attribute assessments relate to the consumer assessments. For example, the expert assessments of 'slippery' and 'washability' will predict the consumer assessments of not greasy, doesn't stain clothes, and washes off easily.

In addition, one will note that 'force to apply' and 'force to spread' assessed by the expert panel will predict 'initial comfort' for the consumer. It is observed that the physical measurement of dynamic friction will predict force to spread which in turn may predict consumer acceptance of 'initial comfort'.



```
EXPERT PANEL                  |              CONSUMER DATA
-----------------------------------------------------------------------------------------
                         - -|
                         - -| ---
Film_ Wash  Init_White   - -|<---------------------------------{L-Value (Chemical)}
             Evenness    - -|
             Slippery    - -| 8     Not_Irritate   Not_Make_Itch
Slip_Wash   Washability  - -| ---   Easy_to_Hold   Fits_Under_Arm     Not_Clothes
            10_Slippery  - -| 7     Dispenses_     Not_Greasy         Washes_Off
                         - -* ---    Controls_Odor  Controls_Wetness  Not_Sticky
                         - -| 6      Cont_WetnessD  Fragrance White_Appl   Odor_As_Need
                         - -| 5      Prot_Stres     Frag_Lasts     Wetness Texture
    Force_Apply          - -| ---
Force_Sprd  Shine        - -| 4      Comfort_Initial<--------{DYNAMIC FRICT.(Physical)}
                         - -|
            15_Rub_Off    0 | 3      Color_Applic  Frag_8Appl    Frag_Men Texture
10Filmy     5Filmy       - -|---     Feel_Appl     Frag_Day15Filmy    5_Rub_Off
                         - -|
            5Coolness    - -| 2
            15_White     - -|
5 _White    Init_Stick   - -+---+    Coolness
                         - -|
                         - -| 1
                         - -|
15_PartRe   White_Wash   - -|
                         - -|
            10_Part_Res  - -|
                         - -|---
            5Part_Res    - -|
                         - -|
         PartResWash     - 1+(0)
-----------------------------------------------------------------------------------------
```

Figure 1; Attribute Map for Expert Panel and Consumer Assessments and Location of Chemical Measurements

The order of importance for acceptance is lack of irritation and lack of itchiness, followed by feel attributes, such as greasiness and stickiness. Next are the performance attributes of controls wetness and controls order. Attributes like coolness and color of the applicator are less important.

CONCLUSION

The Rasch model can provide the tool necessary to combine data from several sources, to relate several kinds of data and clear interpretation of assessments. We also have demonstrated the potential to decrease the number of tests, attributes, and the amount of time and money spent in development.

Objective Measurement of Subjective Well-being

Elizabeth A. Hahn

In everyday situations and during unforeseen circumstances, each of us evaluates the impact of a particular decision in terms of its effect on our quality of life. Although the construct is subjective and is best assessed by self-report, researchers have created acceptable definitions and useful ways to measure it. The following definition is widely accepted for health-related quality of life (HRQOL): "...patients' appraisal of and satisfaction with their current level of functioning as compared to what they perceive to be possible or ideal." (Cella & Cherin, 1988). There are many instruments available to assess HRQOL dimensions such as physical or emotional well-being, as well as disease- or treatment-specific dimensions (Berzon et al., 1995).

Quality of Life in Cancer Treatment

HRQOL is an important consideration in cancer treatment, and healthcare providers seek to improve both the quantity and the quality of their patients' lives. Some cancer types, such as metastatic breast cancer, cannot be cured with currently available therapeutic agents, so the objectives of treatment are directed toward other goals (symptom relief, functional status, prolongation of life). In these patients, the quality of their survival may be as important as the length of their survival. In other types of cancer, the optimal treatment is unknown, and decision-making can best be made by taking into account patient preferences and HRQOL. For example, information about the impact of a disease and its treatment on HRQOL is invaluable for the prostate cancer patient who must decide between 'watchful waiting' vs. surgery, radiation therapy or hormonal therapy, each of which has its own risks and benefits. When treatment costs and health outcomes vary, healthcare providers can use information about preferences and HRQOL to optimize outcomes management.

The focus on HRQOL as an important clinical endpoint in cancer treatment is international in scope. With the availability of multiple language versions of HRQOL instruments, researchers and clinicians are beginning to evaluate the effects of cultural differences on HRQOL measurement. Cross-cultural evaluation of HRQOL and pooling of international research data require unbiased measures of the defined constructs that can detect clinically important differences between patients. Detected differences must not be caused by items that may function differently depending upon patient characteristics.

Elizabeth Hahn

Elizabeth Hahn is a Research Associate with the Institute for Health Services Research and Policy Studies at Northwestern University, and Director of Biostatistics and Data Management Systems at the Center on Outcomes, Research and Education (CORE) at Evanston Northwestern Healthcare. She is a medical sociologist and biostatistician with extensive experience in the design, implementation, coordination and statistical analysis of clinical trials and survey research studies. She also serves as a statistical consultant to international collaborative groups regarding research design and analysis. Her current research includes a focus on methodological and cross-cultural issues in the measurement of health-related quality of life and treatment satisfaction for patients with cancer and other chronic illnesses.

In 1999, she was awarded a two-year grant by the Agency for Healthcare Research and Quality to develop and evaluate a computer-based measurement program for quality of life assessment in low literate cancer patients. She is also the principal investigator on a project to develop a treatment satisfaction scale for cancer, HIV and other chronic illnesses, and a project to evaluate literacy assessment methods and patient preferences and attitudes towards literacy screening.

Cross-Cultural Equivalence

Several types of cross-cultural equivalence have been discussed in the literature, with varying degrees of agreement on definitions and hierarchy (Flaherty et al., 1988; Hui & Triandis, 1985). The universalist approach to cross-cultural research acknowledges that HRQOL concepts may differ across cultures and that this must be evaluated prior to performing comparative analyses. This paper illustrates the use of objective measurement to evaluate item equivalence (commonly defined as items that are relevant and acceptable in both cultures, and that measure the latent trait similarly) and metric/scalar equivalence (the construct is measured on the same metric and locates similar individuals at the same point on the scale).

METHODS

Quality of Life Instruments

The Functional Assessment of Cancer Therapy-Breast (FACT-B; Brady et al., 1997) developed in English, is available in 18 other languages, including German. It includes a general assessment of physical, functional, social/family and emotional well-being as well as a nine-item subscale to assess breast-cancer specific concerns. There are five response categories for the items: "not at all" ("berhaupt nicht" in German), "a little bit" ("ein wenig"), "somewhat" ("m((ig"), "quite a bit" ("ziemlich") and "very much" ("sehr"). The English version of the nine items in the breast cancer subscale are:

I have been short of breath
I worry about the risk of cancer in other family members
I am self-conscious about the way I dress
I worry about the effect of stress on my illness
One or both of my arms are swollen or tender
I am bothered by a change in weight
I feel sexually attractive
I am able to feel like a woman
I am bothered by hair loss

The FACT-B is part of the Functional Assessment of Chronic Illness Therapy (FACIT) quality of life measurement system (Cella, 1997). The initial cultural adaptation of FACIT instruments is based on a sequential approach for the development of internationally applicable quality of life measures, i.e., the instruments are translated from English into other languages (Bullinger et al., 1993). The adaptation methodology involves an iterative forward-backward translation, extensive review and evaluation by bilingual health professionals, and pretesting with patients (Bonomi et al., 1996; Lent et al., 1999).

Patients

The U.S. sample was a subset of 1,616 cancer patients enrolled in a validation study of the FACT-B during 1994-1997. White, English-speaking breast cancer patients (n=195) were selected as a comparison group for the Austrian patients (n=118) who completed the questionnaire in German while receiving treatment at two outpatient clinics during 1995.

Rasch Measurement Model

Rasch (1960) developed the logistic measurement model for the probability of a "correct" response with dichotomous data. This project used an extension of the model for rating scale data i.e., items with ordered response categories such as those used in the FACT-B (Wright & Masters, 1982). The model has three components: 1) an estimate of each patient's "ability" to achieve a high score (high HRQOL), 2) an estimate of each item's "difficulty" (the degree to which an item would be unlikely to be answered in a manner reflecting a high HRQOL) and 3) response "thresholds" for each "step" in the rating scale (there are m-1 steps in an m-category scale). The decisive property of Rasch models is that the person abilities and item difficulties can be estimated independently by means of conditional maximum likelihood estimation, resulting in sample-free question calibration and test-free patient measurement. In the rating scale model, the thresholds can be estimated once for a set of questions.

Item and Metric/Scalar Equivalence

The extent to which items in a questionnaire perform similarly across different reference groups is of critical interest when determining whether a given questionnaire can be used as an unbiased basis for comparing groups. The Rasch model allows us to identify items displaying differential item functioning (DIF). The most important indicator of DIF is not whether items systematically differentiate relevant subgroups, but whether they do so in an unmodeled (i.e., unpredicted) way. Unmodeled differences reflect differential interaction between some items and some persons, which in turn confuses interpretation of results. Items that differentiate groups can be identified and investigated as to their content to determine the likely source of DIF. DIF detecting procedures were applied in four steps: 1) After evaluating and anchoring the step threshold estimates on the entire sample, separate item calibrations were obtained for the two samples. 2) The calibrated item difficulties were plotted against each other. 3) An identity line and statistical control lines (95% confidence limits) were drawn on the plots to guide interpretation and assessment of possible bias (Wright & Masters, 1982). 4) Items identified as possibly biased (displaying DIF) were reviewed to obtain direction on interpreting the plots and determining the appropriate disposition of the item, given the content and the context of the misfitting item. The end product of these analyses and plots is an unbiased subset of items to be used for obtaining patient HRQOL measures on a common, linear metric. The patient measures, rather than raw scores, can then be used for analysis.

The nine breast cancer-specific items in the FACT-

B were evaluated to determine the extent to which they define a unidimensional construct of disease-specific HRQOL. All of the negatively worded items e.g., "I have been short of breath", were reversed in the analyses and item calibrations were reported as logits (log-odd units), with a higher value representing greater item difficulty. The WINSTEPS computer program (Linacre & Wright, 1998) was used to conduct the Rasch model analyses, and SAS software was used to make item difficulty plots.

RESULTS

Patient Characteristics

The majority (57%-60%) of patients (all women) in both groups had no current evidence of disease and few limitations in performance status (81% were classified at the highest level of functioning). The groups were also similar in terms of prior treatment history and current living arrangement. The U.S. group was slightly older and had a higher proportion of patients currently undergoing chemotherapy or receiving hormonal therapy.

Rasch model analyses

Using response thresholds from the combined analysis, separate item calibrations were obtained for the two patient groups and plotted against each other. Only one item ("I am self-conscious about the way I dress") functioned differently across groups. It was more difficult for the Austrian patients. A translation error was discovered in the German language version of this item, which may account for its apparent misfit. The other eight items in the module functioned similarly across groups, suggesting that they can be used to create unbiased measures of HRQOL in Austrian and U.S. breast cancer patients.

DISCUSSION

There is a growing body of literature on cross-cultural evaluation of HRQOL, yet few researchers have appreciated the advantages offered by objective measurement models to control bias and to construct reproducible linear measures. Estimating sample-free item calibrations and test-free person measures provides assurance that the analysis of HRQOL will not be impeded by measurement difficulties.

The limitations of traditional analysis methods to detect bias across different groups of subjects are discussed by Wright, Mead & Draba (1976). Common methods include regression using an external criterion of bias, comparison of factor structures, item-by-group interaction terms in analysis of variance and comparison of the proportion of subjects answering each item correctly. While these methods provide important information about how items function in different groups, they cannot adjust for unequal distributions of person abilities (sample dependency), heterogeneity of item difficulty variance and nonlinearity of raw scores. Rasch measurement model specifies that each item has an inherent property (difficulty level) that does not depend upon any particular sample, and that each person has a characteristic ability (in this case, level of HRQOL) that does not depend upon the particular items used in a test/instrument.

The study reported here demonstrates the usefulness of the Rasch model in evaluating the cross-cultural equivalence of HRQOL instruments. Statistical as well as conceptual criteria were used to determine which items were functioning differently in Austrian and U.S. breast cancer patients. The identification of biased items does not invalidate the questionnaire, but rather enables a better estimate of each cultural group's HRQOL.

References

Berzon R.A., Donnelly M.A., Simpson R.L. Jr., Simeon G.P., & Tilson H.H. Quality of life bibliography and indexes: 1994 update. Qual Life Res. 1995. 4: 547-569.

Bonomi A.E., Cella D.F., Hahn E.A., Bjordal K., Sperner-Unterweger B., Gangeri L., Bergman B., Willems-Groot J., Hanquet P., & Zittoun R. Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. Qual Life Res. 1996. 5, 309-320.

Brady M.J., Cella D.F., Mo F., et al. Reliability and validity of the Functional Assessment of Cancer Therapy-Breast (FACT-B) quality of life instrument. J. Clin Oncol, 1997, 15: 974-986.

Bullinger M., Anderson R., Cella D.,& Aaronson N. Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. Qual Life Res. 1993. 2, 451-459.

Cella D.F. Manual of the Functional Assessment of Chronic Illness Therapy (FACIT Scales) - Version 4. Evanston, IL: Center on Outcomes Research and Education (CORE), Evanston Northwestern Healthcare & Northwestern University, November, 1997.

Cella D.F., & Cherin E.A. Quality of life during and after cancer treatment. Compr Ther. 1988. 4, 69-75.

Flaherty J.A., Gaviria F.M., Pathak D., Mitchell T., Wintrob R., Richman J.A., & Birz S. Developing instruments for cross-cultural psychiatric research. J. Nerv Ment Dis. 1988. 176, 257-263.

Herdman M., Fox-Rusby J., & Badia X. 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. Qual Life Res. 1997. 6, 237-247.

Hui C.H., & Triandis H.C. Measurement in cross-cultural psychology: a review and comparison of strategies. J. Cross-Cultural Psychol. 1985. 16, 131-152.

Lent L., Hahn E., Eremenco S., Webster K., & Cella D. Using cross-cultural input to adapt the Functional Assessment of Chronic Illness Therapy (FACIT) scales, Acta Oncologica, 1999, 38: 695-702.

Linacre J.M., & Wright B.D. A User's Guide to BIGSTEPS/WINSTEPS/MINISTEP: Rasch-Model Computer Programs. Chicago: MESA Press, 1998.

Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogiske Institut, 1960 (Chicago: University of Chicago Press, 1980).

Wright B.D., Masters G.N. Rating Scale Analysis: Rasch Measurement. Chicago: MESA Press, 1982.

Wright B.D., Mead R., & Draba R. Detecting and correcting test item bias with a logistic response model. MESA Research Memorandum Number 22, 1976.

Culture Shift: Managing Change in the Hospital Setting

Judy Schueler and Donna Surges Tatum

Introduction

Since its opening in 1967, the University of Chicago Children's Hospital (UCCH), a 152-operating-bed acute care hospital, has provided comprehensive, innovative medical care to children of all social and economic backgrounds. UCCH is dedicated to preserving the health of children through patient care, education and research into the causes and cures of childhood diseases.

UCCH is staffed by more than 100 physicians of the Department of Pediatrics at the University of Chicago, as well as specially trained nurses and caring support staff, who provide general and specialty medical care for infants, children and teens. The pediatricians of tomorrow - medical students, residents and fellows - also play an important role in caring for children.

The K.I.D.S. First initiative, launched in December, 1997, was designed to fundamentally shift the culture of care at the UCCH. Through interviews and surveys, it was apparent that although staff members were proud to work at UCCH, they believed many barriers existed to delivering optimal care. Additionally, they felt unrecognized for their efforts on behalf of patients and families. It is clear that these perceptions have eroded staff morale and attitudes.

Survey Results

A survey was developed to ascertain attitudes of UCCH staff. The data analysis shows the instrument is well-designed and useful. All of the items fit along the line of inquiry. No items misfit. That is, they are well-written, and are used appropriately by the respondents. They have a reliability of .98. The items are listed in order of how often these behaviors are perceived on the unit. Items above 10.00 indicate a positive response. Those below 10.00 are behaviors that are seen less often. Item maps can be used to devise an Action Plan to improve staff morale and attitudes.

Judy Schueler

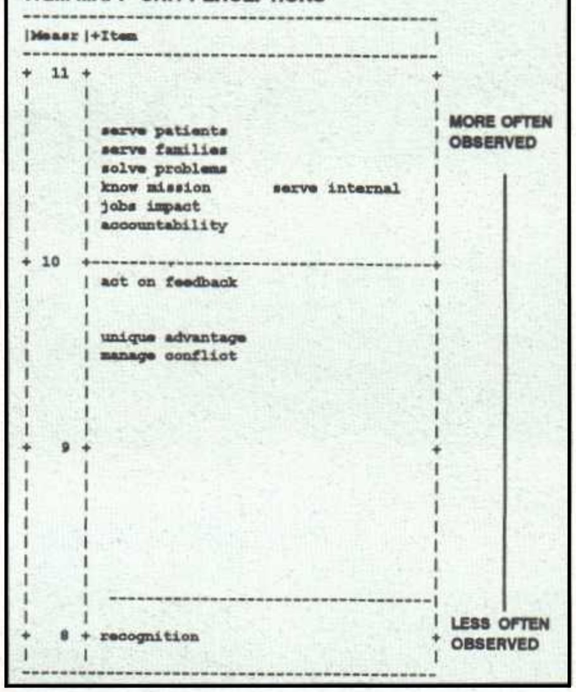
Judy Schueler joined the University of Chicago Hospitals in December, 1992 as the Executive Director of the newly created UCH Academy. Prior to joining the University of Chicago Hospitals, Ms. Schueler served as Vice President of Triton College, River Grove, IL. With over 20 years of experience in curriculum design and higher education, Ms. Schueler has extensive experience in creating School/College/Business partnership programs integrating adult learner services into organizations as well as developing regional retraining assistance centers. The UCH Academy was awarded a "Best Practice" for Education and Training by ASHHRA in 1995. Ms. Schueler graduated with a B.S. in Education as well as a M.S. in Curriculum and Instruction from the University of Illinois. She also posesses a Master's Degree in Management in Organizational Development from Illinois Benedictine College.

Figure 1. Unit Perception

The participants were asked to rate their perceptions of their respective unit and/or department at UCCH on a frequency scale of never; rarely; sometimes; usually; always. The staff perceive themselves to be well-prepared to effectively communicate with and serve patients, families and internal customers. They are encouraged to solve problems and know the mission, vision, direction, and goals of UCCH. They know how their jobs impact patient/customer satisfaction, and employees generally are held accountable for their service-based behaviors and attitudes.

Behaviors which are less often seen are: acting upon feedback; knowing the unique advantages of UCCH over competitors; and knowing specific communication skills for managing conflict. Units are rarely perceived to recognize employees for outstanding service.

FIGURE 1
ITEM MAP: UNIT PERCEPTIONS



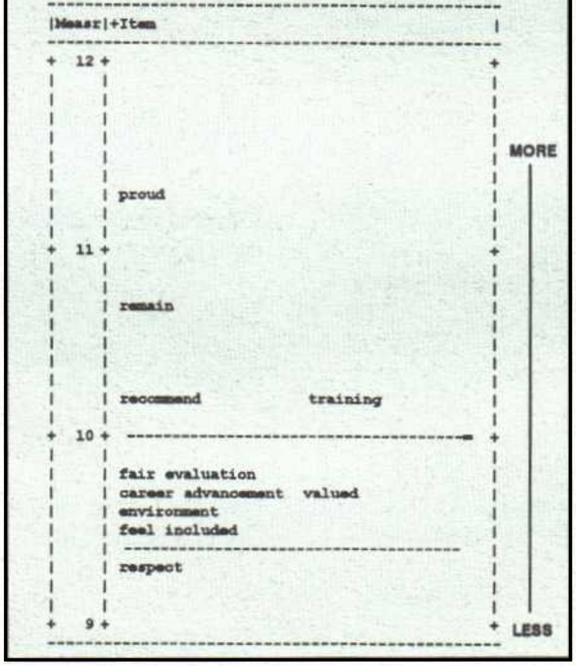
```
------------------------------------------
|Measr |+Item                             |
------------------------------------------
+  11 +                                   +
|     |                                   |
|     |                                   |  MORE OFTEN
|     | serve patients                    |  OBSERVED
|     | serve families                    |
|     | solve problems                    |
|     | know mission       serve internal |
|     | jobs impact                       |
|     | accountability                    |
|     |                                   |
+ 10  +-----------------------------------+
|     | act on feedback                   |
|     |                                   |
|     |                                   |
|     | unique advantage                  |
|     | manage conflict                   |
|     |                                   |
|     |                                   |
|     |                                   |
|     |                                   |
+   9 +                                   +
|     |                                   |
|     |                                   |
|     |                                   |
|     |                                   |
|     |                                   |
|     |                                   |
|     |                                   |
|     |  ---------------------------------|
|     |                                   |  LESS OFTEN
+   8 + recognition                       +  OBSERVED
|     |                                   |
------------------------------------------
```

Figure 2. Personal Perceptions

Participants were asked to rate their perceptions of UCCH from their personal perspective. The rating scale is: not at all; to a slight extent; to a moderate extent; to a great extent; to a very great extent. Only one item slightly misfit: "Do you think you would remain with this organization - even if you were offered a similar job elsewhere?" The responses were a bit erratic on that item, and it did not fit the pattern as well as the rest of the items.

Respondents are overwhelmingly proud to be an employee of the University of Chicago Hospitals. They would remain with the organization even if offered another job; recommend this organization to others; and think opportunities for training are fair and equitable.

FIGURE 2
ITEM MAP: PERSONAL PERCEPTIONS



```
---------------------------------------------
|Measr|+Item                                |
---------------------------------------------
+  12 +                                     +
|     |                                     |
|     |                                     |
|     |                                     |
|     |                                     |  MORE
|     |                                     |
|     | proud                               |
|     |                                     |
|     |                                     |
+  11 +                                     +
|     |                                     |
|     |                                     |
|     | remain                              |
|     |                                     |
|     |                                     |
|     |                                     |
|     |                                     |
|     | recommend            training       |
|     |                                     |
+  10 + ----------------------------------- +
|     |                                     |
|     | fair evaluation                     |
|     | career advancement  valued          |
|     | environment                         |
|     | feel included                       |
|     |  -------------------------------    |
|     | respect                             |
|     |                                     |
|     |                                     |
+   9 +                                     +  LESS
---------------------------------------------
```

Respondents are less sure their performance is evaluated fairly; opportunities for advancement are fair; they are valued; the work environment is supportive and caring; and that they feel included as a member of the UCCH organization. They do not feel all members of the organization are treated with dignity and respect.

Action Plans

There is a renewed focus on enhanced service quality in UCCH. The adaptation of services and programs toward a kid and family orientation recognizes the different and unique needs of children. The K.I.D.S. First initiative aims to incorporate this philosophy into everything that is done at UCCH. The following issues were highlighted during the extensive data-gathering phase. Using the results, many cross-functional teams developed action plans for improving UCCH quality of service. The following has been addressed as the K.I.D.S. First program continues to evolve:

- integrating the UCCH mission into the daily work environment
- developing a pediatric specific candidate assessment program
- creating a pediatric specific interview tool
- implementing a special children's hospital orientation program
- enhancing communication throughout UCCH
- establishing patient satisfaction survey processes throughout the children's hospital
- establishing a reward and recognition program for children's hospital staff
- implementing service improvement initiatives
- measuring the impact of K.I.D.S. First on our patients and staff

Due to the scope and complexity of the K.I.D.S. First initiative, UCCH is interested in determining the impact of the interventions. The collection of baseline data will allow us to subsequently measure our progress and celebrate our successes. Comparative data is scheduled for collection in July of 2000.

Using Rasch Measures For Rasch Model Fit Analysis

George Karabatsos, Ph.D.

In Rasch fit analysis, Z_{ni} is used to measure the fit of a single person-item response, while mean-square (MS) statistics analyze the fit of response sets, and ZSTD tests the significance of a particular MS value.

Most analysts find the Rasch model person measures and item calibrations easier to understand and communicate than the Z_{ni} , MS, and ZSTD statistics. For instance, only through the necessary calculations do we know how much logit-misfit is involved for a given Z_{ni} or MS value. Furthermore, Z_{ni}, MS, and ZSTD are nonlinear functions of Rasch model values (e.g., B_n–D_i).

This paper introduces a Rasch model fit statistic that enables the analyst to interpret fit of a response on the same scale as person measures and item calibrations. Essentially, this is accomplished by explicitly incorporating the logistic Rasch model in the fit statistics.

RESPONSE-FIT INDEX FOR DICHOTOMOUS CHOICES

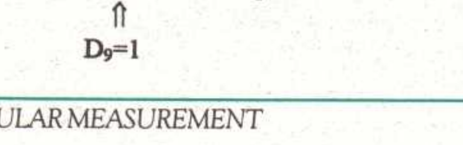
Let K_{ni} denote the logit-fit of person n's response to item i, calculated by: $K_{ni} = f_{ni}(B_n - D_i)$ [1]

where f_{ni} classifies the model-fit of a person-item response

$f_{ni} = 0$ for a response that fits the model
($X_{ni}=1$ when $B_n \geq D_i$, or $X_{ni}=0$ when $B_n \leq D_i$)
$f_{ni} = -1$ for a response that misfits the model
($X_{ni}=1$ when $B_n < D_i$, or $X_{ni}=0$ when $B_n > D_i$).

Example 1. *Richard with ability $B_R=3$ encounters "item 9" having difficulty $D_9=1$.*

Map item and person on a number line:

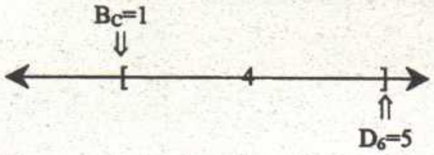


Expected Response Rule: Since $B_n > D_i$, then $\{X_{ni}=1\}$ is the expected response.

<u>Two Possible Scenarios:</u>

Response	Fit result	Interpretation
$\{X_{ni}=1\}$	$K_n = 0(3-1) = 0$	Response fits measurement model.
$\{X_{ni}=0\}$	$K_n = -1(3-1) = -2$	Richard responded 2 logits *below* expectation.

Example 2. Cindy with ability $B_C=1$ encounters "item 6" having difficulty $D_6=5$.

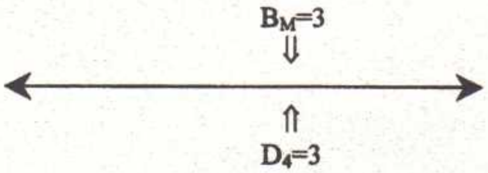


<u>Expected Response Rule:</u> Since $B_n < D_i$, then $\{X_{ni}=0\}$ is the expected response.

<u>Two Possible Scenarios:</u>

Response	Fit result	Interpretation
$\{X_{ni}=1\}$	$K_{ni} = -1(1-5) = 4$	Cindy responded 4 logits above expectation.
$\{X_{ni}=0\}$	$K_{ni} = 0(1-5) = 0$	Response fits measurement model.

Example 3. Mary with ability $B_M=3$ encounters "item 4" having difficulty $D_4=3$.



<u>Expected Response Rule:</u> Since $B_n=D_i$, then $\{X_{ni}=0\}$ and $\{X_{ni}=1\}$have equal probability ($P_{ni1}=.50$, therefore $P_{ni0}=.50$). So by definition, *neither response misfits the model.*

Two Possible Scenarios:

Response	Fit result	Interpretation
$\{X_{ni}=1\}$	$K_{ni} = 0(3\text{–}3) = 0$	Response fits measurement model.
$\{X_{ni}=0\}$	$K_{ni} = 0(3\text{–}3) = 0$	Response fits measurement model.

RESPONSE-FIT INDEX FOR POLYTOMOUS CHOICES

Since all Rasch models reduce to the dichotomous-response model, Equation 1 can be extended to analyze the fit of a rating-scale response. For an item with m response categories, there are m–1 adjacent-category steps, where each step j is denoted by the parameter F_j. A person's rating scale response to that item indicates a certain number of "advanced" steps, and a certain number of "unadvanced" steps. Each "advanced" versus "unadvanced" step response is a dichotomy, and therefore, there are j dichotomous responses within a single rating scale response.

The fit calculation of a single rating scale response involves calculating $f_{ni}(B_n\text{–}D_i\text{–}F_j)$ for each of the steps, and letting K_{ni} equal the calculation that differs the most from zero. The K_{ni} for a single rating scale response is therefore calculated by:

$$K_{ni} = |\max|\ [f_{nij}\ (B_n\text{–}D_i\text{–}F_j\ )] \qquad [2]$$

where,

$|\max|$ maximum in absolute value

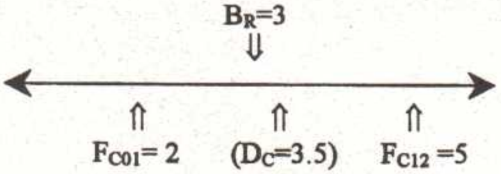
$f_{nij} = 0$ for a step-response that fits the model

$f_{nij} = -1$ for a step-response that misfits the model

In the case of dichotomous response choices, there is only one threshold j, in which case equation [2] reduces to equation [1].

Here is an example of an item with a three category (m=3) rating scale, where $X_{ni}=\{0,1,2\}$, rendering m–1=2 steps. Let F_{01} denote the parameter for the step to category 1 from 0, and F_{12} for the step to category 2 from 1.

Example 4. Bob with ability $B_B=3$ encounters "item C" having difficulty $D_C=3.5$, where $F_{01}=?1.5$ and $F_{12} = +1.5$ relative to D_C.



Expected Response Rule: Since $B_n>F_{i01}$ and $B_n<F_{i12}$, , $\{X_{ni}=1\}$ is the expected response.

Three Possible Scenarios:

Response	Fit result	Interpretation
$\{X_{ni}=2\}$	$K_{ni}= |\max|\ [0(3\text{–}2), -1(3\text{–}5)] = 2$	Bob responded 2 logits above expectation.
$\{X_{ni}=1\}$	$K_{ni}= |\max|\ [0(3\text{–}2), 0(3\text{–}5)] = 0$	Response fits measurement model.
$\{X_{ni}=0\}$	$K_{ni}= |\max|\ [-1(3\text{–}2), 0(3\text{–}5)] = -1$	Bob responded 1 logit below expectation.

FIT ANALYSIS OF RESPONSE SETS

Analyzing response sets is straightforward. The average of the absolute value of $|K_{ni}|$ values can be taken across all responses of interest:

$$|\overline{\mathbf{K}}_{\mathbf{ni}}| = \frac{\sum |\mathbf{K}_{\mathbf{ni}}|}{\mathbf{N}_{\{\mathbf{X}_{\mathbf{ni}}=\mathbf{x}\}}} \qquad [3]$$

to obtain the "average logit noise," where $N_{\{Xni = x\}}$ denotes the total number of responses. Person $|\overline{K}_{ni}|$ is obtained by applying Equation 3 for all person responses; item $|\overline{K}_{ni}|$ is calculated for all item responses.

It is also informative to take the average of certain response subsets. Examples include (1) the subset of "negative" K_{ni} values, and (2) the subset of "positive" K_{ni} values. Subset (1) indicates the magnitude of surprising "low" responses (e.g., occurring from sleeping, carelessness, etc.), and subset (2) indicates the magnitude of surprising "high" responses (e.g., lucky-guessing).

The accuracy of K_{ni} depends on parameter values estimated from the data, *but we know we estimate parameters from noisy data in the first place* (*Z_{ni}, MS, ZSTD*, and all parameter-dependent fit methods suffer this uncertainty). When data noise is high, we cannot trust the accuracy of parameter estimates, and therefore can no longer trust the accuracy of K_{ni} and other parameter-dependent fit statistics. In cases where data is too noisy for the parameter-dependent fit statistics to be useful, an alternative is a an estimate of Guttman fit:

$$G = \frac{N_{|K_{ni}|>0}}{N_{\{X_{ni}=x\}}} \qquad [4]$$

which is the proportion of unexpected responses across the relevant response set. G is linearized by the transformation log(G/(1–G).

It is also informative to change the numerator of Equation [4] to calculate the proportion of surprising "low" responses ($N_{K<0}$) and "high" responses ($N_{K>0}$).

G interprets Kni values as ordinal (possible values: either $K_{ni}=0$ or $|K_{ni}|>0$), which renders it more robust than $|\overline{K}_{ni}|$ (and Z_{ni}, MS, ZSTD) to inaccurate parameter estimations. Hence, G can be considered a parameter-free fit statistic.

CLASSROOM CLASSICS

From the Classroom

Classic Instructional Handouts of Professor Ben Wright

Anyone who has taken courses with Professor Ben Wright at The University of Chicago probably still keeps a treasured collection of Ben's famed class handouts. Ben has a genius for moving his provocative ideas from his mind to ours.

He swears that they begin in the pool, where he works out solutions to intellectual problems during his early morning swims. Then they take shape in rapidly created words and pictures that gradually fill the blackboards in front of fascinated students in Judd Hall. Finally, when Ben is satisfied that he has an idea firmly in his sights, he designs bold and provocative printed handouts for students to return to and ponder again and again. It occurs to us that these wonderful pedagogical materials deserve a wider audience.

Beginning with this issue, *Popular Measurement* will begin to reprint the best of Ben's handouts. We hope they will delight former students and intrigue interested readers who are not yet acquainted with them. Reactions and requests for old favorites are welcome.

Matthew Enos

What's to Learn in Psychometrics?

Ben Wright

I. BASICS

- A. The only theory useful to you is one you know well enough to invent and verify
- B. The distinction between **quantitative** differences of degree and **qualitative** differences of kind
- C. The necessity and opportunity for social science to be **as quantitative as physics**
- D. A **useful variable** is a workable fiction indicating quantities of **one and only one** thing
- E. For a measure to have **meaning**, its line of increase must be **benchmarked** by **calibrated explanatory item content**
- F. How to **construct useful measurement** from **ordered nominal observations**

II. MEASUREMENT

- A. Observations must be **replicated** to accumulate and focus the information they are intended to imply
- B. **Counts** of replicating observations are the **scores** necessary to construct measures
- C. Scores must be statistically **sufficient** for measurement to occur
- D. But **scores are not measures** because:
  - 1. Scores are ordinal - not linear (additive)
  - 2. Scores are test and sample dependent - **not objective**
  - 3. Scores, on their own, **cannot be validated**
- E. Measures, in contrast to scores, are:
  - 1. **Additive**, linear, interval
  - 2. **Objective**, invariant, generalizable
  - 3. **Error** qualified for their estimation unreliability
  - 4. **Fit** validated for their one dimensional coherence
- F. When the score-to-measure function necessary to satisfy any reasonable measurement requirement is deduced, the Rasch model is found to be the necessary and sufficient result - this means that:
  - 1. Fit to the Rasch model is the necessary and sufficient condition for constructing measurement from data
  - 2. Only data which can be made to fit the Rasch model can be useful for constructing measurement

III. STATISTICS

- A. Are **never** perfectly **reliable**
  - 1. Their inherent **error must be estimated** and reported
  - 2. Inferences about measure distributions and regressions will be mistaken unless their **statistics are corrected for measurement error**
- B. Are **never** completely **valid**
  - 1. The extent of invalidity must be assessed, allowed for in estimation error and reported
  - 2. Improbable data signifying qualitative differences must be detected, identified, diagnosed, isolated and reported
- C. Always require **visualization:** graphing, plotting, mapping for **comprehension** and **communication**

Three "Cs" to Meaning: The Big Picture

Ben Wright

CONSTRUCT

1. Intention / hierarchy / dimension / variable
   leading to a **Construct MAP**
2. Realization / articulation / itemization / ITEMS / item positioning
   leading to a **Questionnaire**

CONVERSATION

1. Invitation / motivation / convenience / comfort / security
2. Linguistics / language verification
3. Response format
   post-code / precode
   circle / check / fill

MEDIA FOR CONVERSING

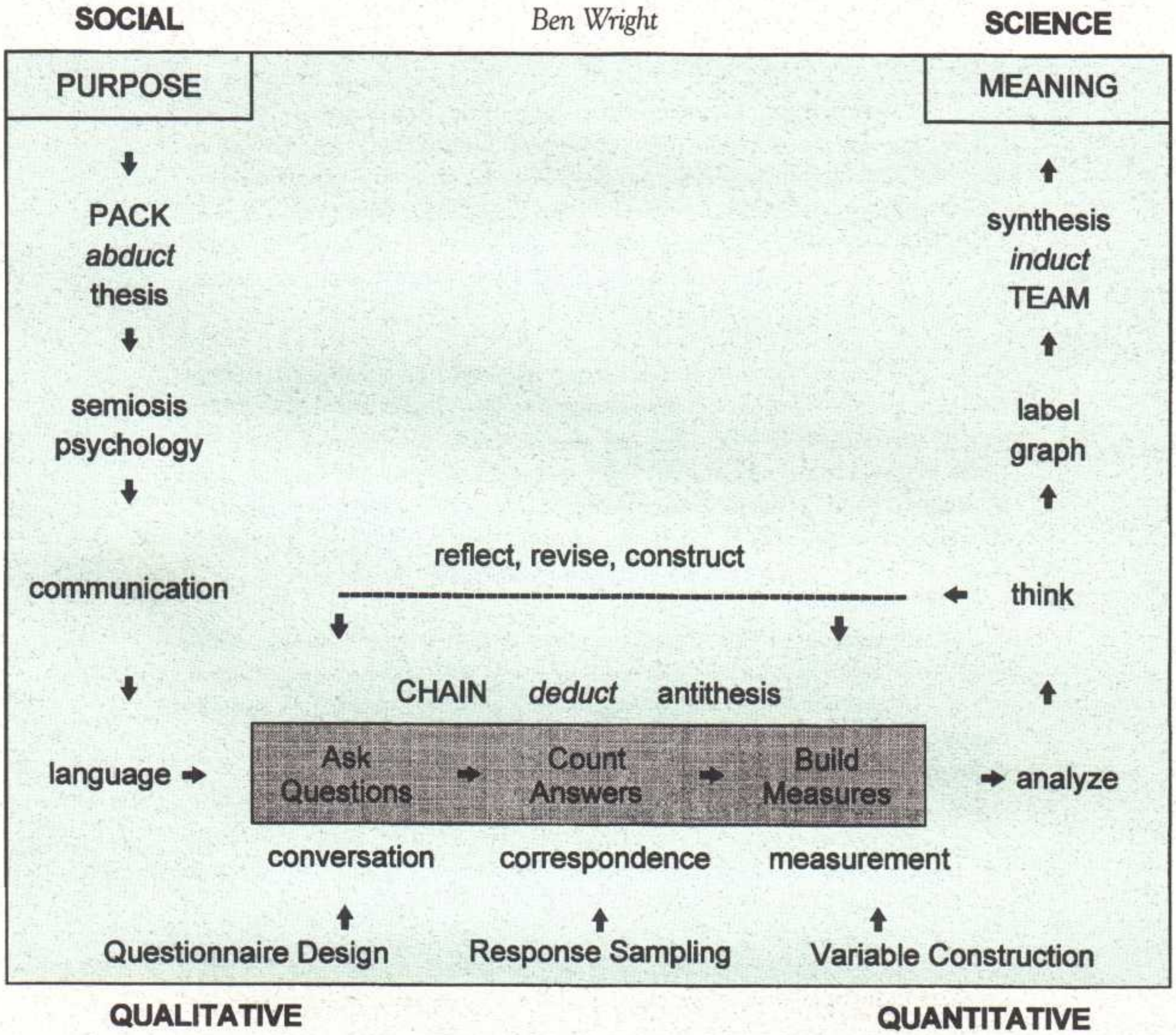
Opinion:	agree / disagree	Value:	good / bad
Attitude:	like / dislike	Behavior:	do / don't
Frequency:	often / seldom	Amount:	a lot / a little
Force:	strongly / weakly	Involvement:	actively / passively

COMPREHENSION

1. Scoring model
2. Measurement model
3. Item analysis, diagnosis, revision -
   leading to a **Construct** (Criterion) **MAP**
4. Person analysis, diagnosis, editing -
   leading to an **Application** (Normative) **MAP**

The Road to Reason

Ben Wright



There are no pat answers on the road to reason, but there are many satisfying questions. We start from deep within ourselves, with Peirce's signs (*RMT, 11:1,* 539-540). Our brain cells work as a pack of hounds each searching for the prey (*RMT, 10:2,* 501). We abduct in thought, making intuitive leaps, defying logic, as we strive to formulate ideas expressible as words in some thesis. Then we communicate it to ourselves and others, searching for qualitative instances of what might be it.

There is no contradiction or conflict between the qualitative and the quantitative. The qualitative is complex, inscrutable, unique. But to learn from it, utilize it, manipulate it, it must be made simple, obvious, general. The leap from qualitative to quantitative is based on this organizing principle. We want to escape the contradiction, chaos and idiosyncrasy of the impractical concrete. We want to build a manageable "world" based on the practical abstract.

Rasch measurement is our construction tool. In a careful process of deduction, we pile up the qualitative. We compress it. We chip off protuberances, smooth off rough edges to arrive at an artifact as elegant and handcrafted as ever formed from raw material by inspired craftsman.

But does our artifact have value? Is it a bauble or a gem? We must think. We must analyze. We must induce what greater meaning our artifact embodies. This prompts speculation, new abduction, and we're off to the beginning of a further road to reason.

Rasch Measurement Transactions, 11:4, 589 [rev]

CLASSROOM CLASSICS

Realizations of Measurement

Ben Wright

Every morning I squeeze the orange juice. Two glassfuls - one for Claire, one for me. Oranges are very much themselves, each orange an individual in size, color, fragrance, softness. I can count oranges. How many shall I get out of the icebox to fill the two glasses? The trouble is that there is no constant number of oranges that makes a glassful. Sometimes it's only three. Other times it can take six. How on earth can I regularize this procedure?

Well, as I am sure you've already guessed, the solution is embarrassingly simple. The Co-op sells oranges in four-pound bags. No matter how many oranges it takes, the bag always weighs four pounds. Now "A pint's a pound the world around," and an orange is about one-fourth juice, by weight. By experiment and calculation, I establish that two pounds of oranges makes a glassful - no more, no less. This is always so, no matter how few or how many oranges it takes to weigh the two pounds.

The result of this abstract science is that I have a simple, foolproof, infinitely reproducible, inferentially stable rule. Take one four-pound bag of oranges out of the icebox and squeeze whatever number of oranges happen to be in it. Two glassfuls are always produced - no matter how many or how few the oranges.

Weighing oranges is vastly superior to counting them - perhaps not for art or even literature but certainly for routinely obtaining a glassful of orange juice.

But counting each orange is so immediate, so fulsome, so personal, so individually appreciative, and so richly qualitative. While weighing bags of oranges is so impersonal, so meagerly singular, so general, and so unappreciative of the truly unique, individual nature of each orange, so niggardly quantitative. How dare I reduce the lovely, charming, richly multidimensional orange to a mere cold, stingy weight in lifeless, uncaring pounds. What a travesty of nature!

But what a triumph for obtaining a glassful of orange juice every time. You might decry my reduction of the gorgeous orange to such a brutal simplicity as its weight. But you have to admit that for routinizing the production of glassfuls of orange juice you will never in a million years invent an approach that is as simple or as reliable. That's the difference between art and science, between counting right answers and constructing measures.

Basic Research Methods

Ben Wright

A. Five Psychological Data Construction Procedures

1. The BEHAVIOR. The show! What you perceive: see, hear, smell, feel, taste. What the person manifests.
2. The EFFECT. Your response! What the person *does to you*. Your experience as the *object* of their behavior.
3. The FEELING. Empathy, identification! Who *you become* when you are a *subject* behaving their way.
4. The CLAIM. The story! What the person *says they're doing.*
5. The INFERENCE. What *you deduce* from theory to be the meaning which follows from any or all of the above.

B. Four Research Design Principles

1. IDENTIFYING categories: *naming.*
2. REPLICATING identities: *counting.*
3. CONTROLLING identifiable interactions and interferences: *matching, blocking, stratifying.*
4. RANDOMIZING unidentifiable interferences: *sampling, assigning, distributing.*

C. Three Measurement Requirements

1. UNITS to count with: linearity, additivity, *differences.*
2. ORIGINS to count from: multiplicativity, *ratios.*
3. INVARIANCE to count on: objectivity, *generality.*

D. Three Statistical Requirements

1. AMOUNT: **measure** estimated through a measurement model.
2. ACCURACY: **error** of estimation defined by the measurement model; precision, margin of error, *reliability.*
3. COHERENCE: **fit** of these data to the measurement model; consistency, data quality, *validity.*

CLASSROOM CLASSICS