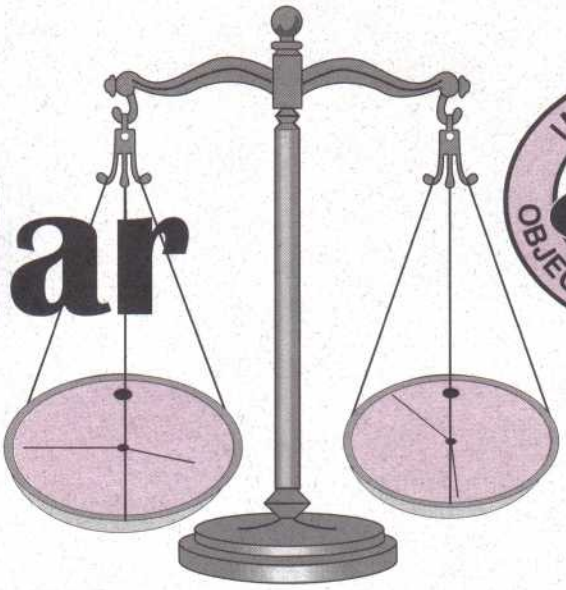
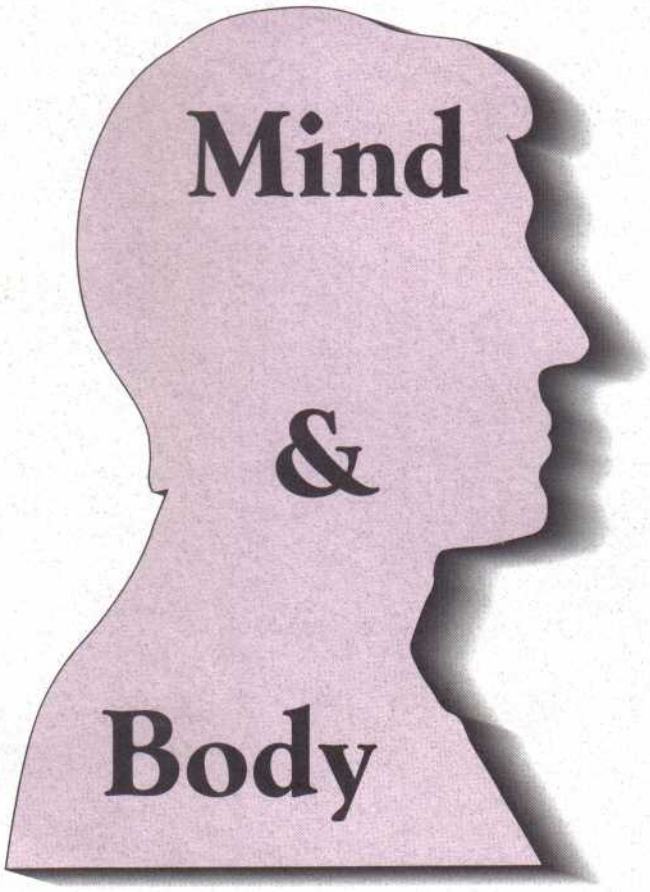
Volume 2 - No. 1 - $10.00
Spring 1999

Popular Measurement



INSTITUTE FOR OBJECTIVE MEASUREMENT
1996
IOM

Journal of the Institute for Objective Measurement



Inside:

INDEX

NEWTON: PINBALL WIZARD?

Larry H. Ludlow[1]
Kathleen A. Haley
Boston College

In 1981 the first author worked as Geoff Master's MESA Psychometric Laboratory assistant. His task was to write code for the partial credit Rasch model. While testing the program a computational problem was encountered that did not make sense, regardless of how, or how often, Geoff or Ben tried to explain it. The problem was that some of the calculations were shrinking to zero and the program was crashing. Geoff and Ben explained this phenomenon as a failure to converge on the part of the Newton-Raphson technique. That explanation did not mean much to him then nor does it mean much now to many of our beginning psychometrics students.

In these days of specialization, it is amazing and humbling to reflect on the great number of areas in which Newton not only had interest, but great influence. Optics, astronomy, chemistry, physics, and mathematics are notable examples. An Alexander Pope couplet is often quoted to demonstrate Newton's influence in his lifetime:

Nature, and Nature's Laws lay hid in Night.
God said, Let Newton be! and All was Light.

Newton's beginnings, however, were not auspicious. He was born dangerously prematurely and was lucky to survive. His father died before he was born, and his mother was absent for much of his early childhood, having left him with his grandmother when she remarried (Christianson, 1984).

Although his work changed science in ways we rarely think about, one of his greatest contributions to our field was in pure mathematics. His work in the invention of differential calculus allows us to find minima and maxima of curves, without which we would be missing many standard statistical techniques. In addition, while trying to solve Kepler's equation for the position of a planet at any given time he developed a numerical procedure for solving higher-order polynomials that could not be solved using calculus (Pepper, 1988).

The rationale behind the technique is relatively simple. That is, if we don't know how to obtain a direct estimate of a parameter, then use a rough initial estimate that we can iteratively adjust. Different numerical analysis strategies exist for obtaining initial estimates, their subsequent adjustments, and the final estimates. Newton's contribution was that he developed the first practical technique for solving such a problem. The technical detail is standard material in numerical analysis texts and is included in the appendix.

The way the original programming problem came to be understood, however, was by printing out all the intermediate values of the calculations and then plotting them.

Figure1

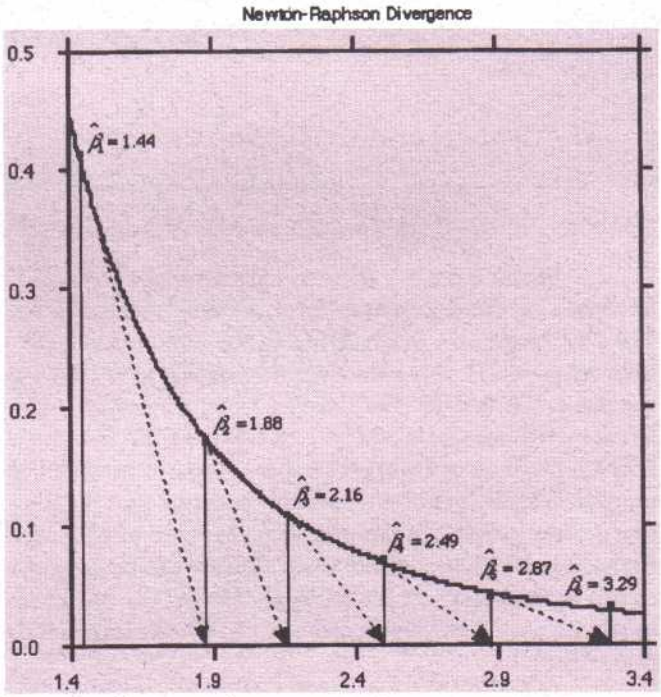


Figure 1 illustrates the situation that prompted the study of the Newton-Raphson technique. In this situation we see divergence, or a failure to converge. The initial, intermediate, and final calculations **(see Table 1)** in the iterative process to arrive at the logit ability estimate for an extremely high raw score are plotted. The estimates ($\hat{\beta}$) become more extreme as the part of the equation that reflects the precision of the estimates

(DB) goes to zero causing the actual adjustment (DINC) to increase without limit. This figure did not offer particularly useful insight into the technique, however, because it does not show what a proper solution should look like. (NOTE: the terms GS, DB, and DINC were variable names originally used in the programs SCALE and CREDIT and may very well still be used in whatever Rasch software you are currently using. In addition, most programs have built-in checks that slow down DB from going to zero—the "1/2 Correction" factor applied to the adjustment.)

Figure 2
Newton-Raphson Convergence

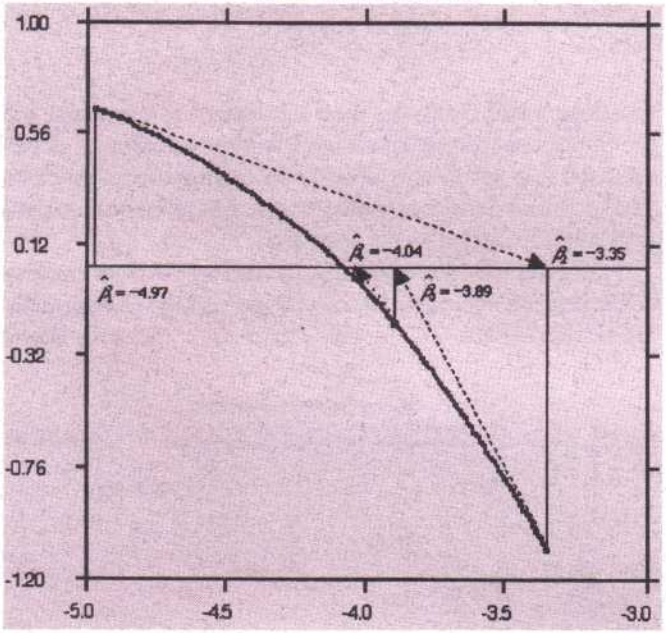


Figure 2 is a plot of the calculations **(see Table 2)** to arrive at a logit ability estimate for an extremely low raw score. The plot illustrates a convergence in the estimates. The first estimate is (-4.97). The second estimate swings to the right on the line and is (-3.35). The third estimate swings back to the left along the line and is (-3.89). Note also how the adjustment (DINC) becomes increasingly smaller. After studying plot after plot of similar patterns it suddenly occurred that the iterative process (when it converges) is analogous to a ball bouncing back and forth in a pinball game until the ball finally comes to rest at the bottom of the machine. Hence, the title of the article (although converging estimates do not necessarily have to oscillate).

Ironically, Newton would not likely have found humor in the analogy or the title of this article. In fact, he was described as "anything but affable." He was extremely unwilling to give credit to others or share credit for great discoveries. For example, he conducted a near-lifelong priority dispute with Leibniz over the invention of calculus, and he had a long-standing dispute with Johann Bernoulli. He was acrimonious in dispute, vindictive, imperious, insulting, and arrogant. His "academic overkill" even went so far as to lead him to brag "He had broke Leibniz's Heart with his Reply to him" (Boorstein, 1983).

Finally, how did "Raphson" become associated with the technique? John Raphson was an accomplished mathematician in his own right, having been elected to the Royal Society the year *before* his graduation from Jesus College Cambridge. He published a mathematical dictionary and *De spatio reali*, an application of mathematical reasoning to theological issues. He also published another theological book, *Demonstratio de deo* (O'Connor & Robertson, 1998). However, he is better known by his association with Newton. Quite simply, the numerical technique developed by Newton in 1671 came to be called the Newton-Raphson technique because it was first published in Joseph Raphson's 1690 *Analysis aequationum universalis*. Newton himself did not publish it until 1736 (O'Connor & Robertson, 1998).

References

Boorstein, D.J. (1983). The Discoverers. *New York: Random House.*

Christianson, G. (1984). In the Presence of the Creator: Isaac Newton and His Times. *New York: The Free Press, Macmillan.*

Pepper, J. (1988). Newton's mathematical work. In Fauvel, J., Flood, R., Shortland, M., and Wilson, R. (eds.), Let Newton Be!. *Oxford: Oxford University Press.*

Appendix

The pinball analogy often works better in our courses than the traditional explanation found in numerical analysis texts: Given a function f(x), solve for the roots of f(x) such that f(x)=0. To attempt to find that location where the roots of the function are 0 requires some initial estimate for the location f(x)=0. Once an initial estimate is provided we try to improve upon it. The solution is based on the observation that if x_0 is close to a 0 of f(x), then the tangent to the graph of f(x) at $(x_0, f(x_0))$ intersects the axis X at a point, say $x^{(1)}$, which is closer than x_0 to the 0 of f. The process then consists of computing $x^{(1)}$, substituting it back into the equation as x_0, and recomputing $x^{(1)}$ until some level of desired accuracy is achieved. This process is the Newton-Raphson method of solving for the roots of an equation.

For our purposes the important piece in this process becomes $x^{(1)} = x_0 - \frac{f(x_0)}{f'(x_0)}$. The numerator (called GS) gives the direction in which the adjustment should be made and the denominator (called DB) reflects the magnitude of adjustment (DINC). Technically, the second derivative of the function must not go to zero (shown by the asymptotic function) which is exactly what happens with some data sets containing extremely high or low person or item scores.

In practical terms, when we are estimating β for persons the function f(x) we are solving (GS) is the difference $(r - \sum p)$ where r is a person total score and $\sum p$ corresponds to the estimated total score. The value $x^{(1)}$ corresponds to our new estimate of $\hat{\beta}$ and x_0 corresponds to our previous estimate of $\hat{\beta}$ (which on the first iteration came from PROX).

Table 1. Divergence

Iteration	$\hat{\beta}$	GS	DB	DINC	1/2Correction
1	1.44	.41	-.94	.43	
2	1.88	.17	-.3	.57	.28
3	2.16	.11	-.16	.67	.33
4	2.49	.07	-.09	.76	.38
5	2.87	.04	-.05	.83	.42
6	3.29	.03	-.03	1.86	.43

The following example shows how the equation works:

Given that $x^{(1)} = x_0 - \frac{f(x_0)}{f'(x_0)}$, then $\hat{\beta}_2 = \hat{\beta}_1 - \frac{GS}{DB}$, or

$1.88 = 1.44 - \frac{.41}{-.94}$ where $DINC = -\frac{GS}{DB}$ or .43.

Table 2. Convergence

Iteration	$\hat{\beta}$	GS	DB	DINC
1	-4.97	.65	-.4	1.62
2	-3.35	-1.09	-2.03	-.53
3	-3.89	-.2	-1.31	-.15
4	-4.04	-.01	-1.12	-.01

[1] This paper is distilled from one originally written in 1981 with the statistical assistance of Graham Douglas and Geofferey Masters. The original paper is still used in our courses.

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)

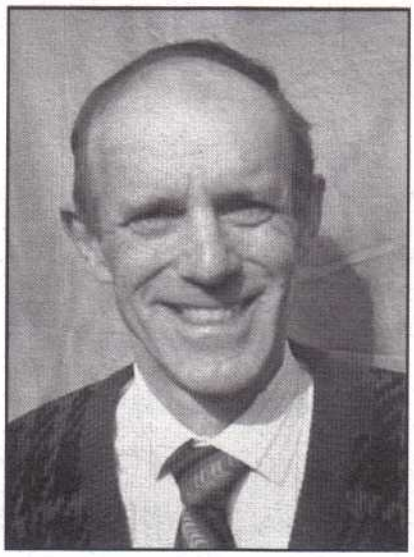
e-mail: LUDLOW@BC.EDU

John Michael Linacre

The Man Behind the Code

Ellen Sullivan Woods

Johann Wolfgang von Goethe's words of advice for the apt educator ring eminently true for Michael Linacre, "Only by errors that irk us do we advance." Also, equally befitting Mike is Goethe's definition of the educator's task ... "not to preserve from error but to guide the errant, indeed to let them savor their errors to the dregs——such is the teachers's wisdom. "Whoever barely tastes his error will long nurse it, will revel in it as though in a rare treat; but whoever drains it to the bottom must come to know it." Indeed, the heart of Goethe's quote resonates for any student of Mike's. The lingering effect of his wonderful crisp presentations and effulgent ideas produces a curiosity that is sadly rare for many students of statistics.



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

Measurement classes at the University of Chicago are largely comprised of individuals that have used statistical measurement techniques throughout their careers but have never quite known the richness of the typical findings that come out of this class. Dr. Linacre delineates new ways to look at the problem of research that both enhance scientific progress as well as the researchers use of a variety of theoretical underpinnings. All students' work bears the same weight in the eyes of Dr. Linacre. "It is our duty to tell others but what they are able to receive. Man grasps but what is his measure," reminds Goethe.

Indeed, there is but one Michael Linacre but with many facets. (This is no coincidence since Facets is the name of the new phase he launched in 1989 of the Rasch software program that he designed in cooperation with the American Society of Clinical Pathologists.) This innovative program, while implementing Rasch analysis, provides a unique focus intended to construct measures from observations based on the expert rating of examinee performance.

Dr. Linacre needs little introduction to members of the field of measurement. Mention must be made of the year 1989, however, since it was important for both Dr. Linacre, Maureen, his wife, and certainly for Chicago. In that year he obtained a Ph.D. in Psychological Measurement from the University of Chicago and took up research and teaching there as well. One of Mike's many notable talents is his ability to transfer proven methods to his classroom audience. Anyone within his reach feels that he can tap into Mike's genius for skillfully linking measurement ideas with the larger scheme of things. His answers invite questions. When a classroom is exposed to Dr. Linacre's spirited ideas, his easy laugh and engaging smile inspire enthusiasm for one of the most understood and oft-maligned disciplines——measurement. He is a master teacher consistently confuting the monody of the "expert assumption" that postulates that "effectiveness can't be taught."

Mike was born in England. In 1967, he graduated from Cambridge University with a BA in Oriental Studies and Mathematics. In 1971 he was awarded an MA by Cambridge University. Also, in 1971 he obtained a second BA degree from Ambassador College (Bricket Wood England). In subsequent years, he obtained further degrees and certifications in Psychological measurement together with a second MA from the University of Chicago. In 1989, he decided to take up research and teaching at the University.

Fascinated by computers as a young adult, Mike became involved with various programming projects throughout college even while attending Cambridge as an undergraduate. He engaged in computer-related technical programming and management activities in England, Japan, Australia, and the USA, and by 1981 he was Computer Sciences Manager on a project to develop test instruments to evaluate local HEAD START programs for the Administration for Children, Youth and Family (ACFY) of the Federal Department of Health and Human Services (HHS). It was at this time that a very important friendship developed with Dr. Benjamin D. Wright. By 1986, due to the inducement of Dr. Wright, Mike and his wife, Maureen, moved to Chicago and to the University of Chicago. After receiving a Spencer Dissertation Year Fellowship, he ob-

tained a Ph.D. in psychometrics and educational measurement under the aegis of Dr. Wright, a pioneer in the practical application of Rasch measurement from many different types of social science data. (If greatness comes to those who team up to make cooperation, vision, and personal understanding a reality, then the alliance he and Dr. Wright have forged over the past 10 years is a testimonial to teamwork of the highest degree.) By 1989, MESA's previous computer programs for multiple choice tests, attitude surveys, and partial credit response formats (BICAL, CREDIT, MSCALE and MSTEPS), were superseded by BIGSCALE, introduced in 1991, and WINSTEPS in 1998, gave Dr. Linacre extensive knowledge in the 'anatomy of assessment.' As the Associate Director of the Measurement, Evaluation, and Statistical Analysis (MESA) Psychometric Laboratory at the University of Chicago, Mike is very active in the application and dissemination of Rasch analysis techniques. He is the editor of the Transactions of the Rasch Measurement Special Interest Group (SIG) of the American Educational Research Association and also SIG operations manager. While he lectures and publishes widely, Mike also devotes time to his responsibilities as pastor of the Active Bible Church of God since it foundry in 1996 in Hyde Park, a neighborhood of Chicago. Mike also is a pastor of the Biblical Church of God in Danville, Illinois and the head coordinator of the Student Christian University Bible Association, a campus outreach at the University of Chicago. Mike is a consultant to major public and private testing agencies. He admits that the primary challenge now and in the future facing practitioners of the Rasch methodology is improvement of the communication of findings to the decision-makers and, more importantly, to society at large.

In the light of his own stated challenge to explain testing results more effectively, it is a great consolation to his colleagues that much of Mike's success to date rests on his consummate ability to translate technical visions into a marriage of strategy for both the market and the academic place. He blends creativity and technical acumen in an alchemy that distinguishes Mike in the classroom, in his research, and the marketplace.

Ungar, Frederick. 1989. Goethe's World View: Presented In His Reflections And Maxims, Frederick Ungar Publishing Company. New York, New York. Linacre, John Michael. 1998, Spring. Ben Wright: the Measure of the Man, Popular Measurement: Journal for The Institute for Objective Measurement.

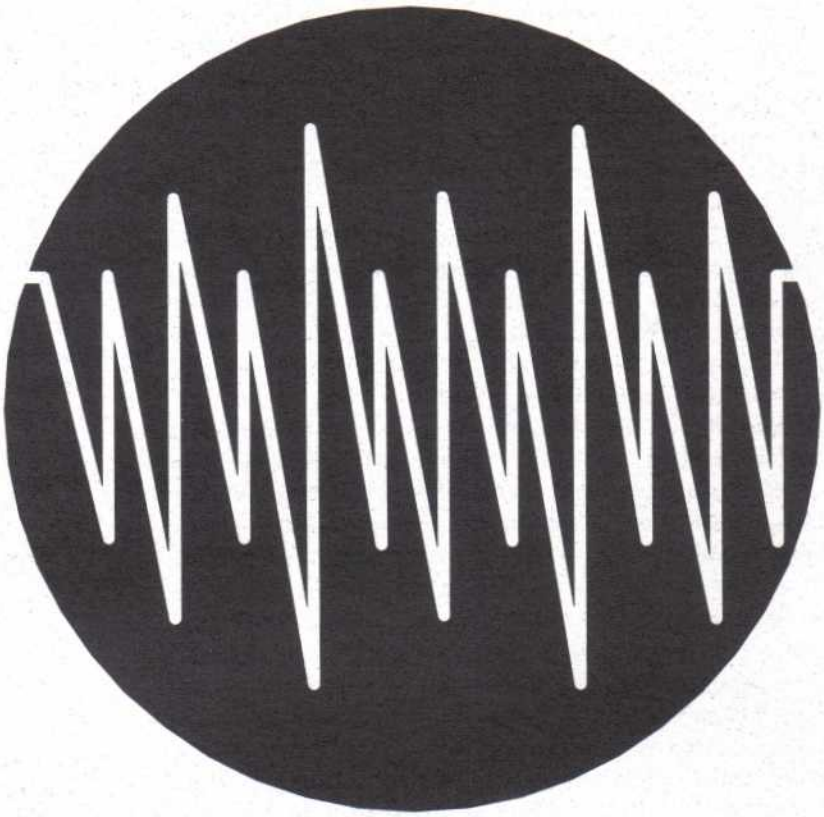
Ellen Sullivan Woods

Ellen Sullivan Woods is a recent graduate of the Masters In The Science of Communications program at Northwestern University. She is a wife and mother of 5 grown children ages 15 to 30. She serves as: Commission member of Animal and Care Control for the city of Chicago; Lector, St. John Fisher Parish; Leadership Committee, Alivio Medical Center; Education Committee, Executive Club of Chicago; Advisory Board Member, Black*Star Project. She loves to teach watercolors and gardening. In her leisure time, she enjoys travel. Ellen looks forward to further research and involvement in the use of psychometrics in its application to speech and communication training.

"Our analysis which has hitherto been qualitative, must become quantitative; we must cease to be empirical, and become scientific: in criticism as in other matters, the test that decides between science and empiricism is this: `Can you say, not only of what kind, but how much?' If you cannot weigh, measure, number your results, however you may be convinced yourself, you must not hope to convince others, or claim the position of an investigator; you are merely a guesser, a propounder of hypotheses."

Frederick Gard Fleay (1831-1909), British Shakespearian scholar in
On metrical tests as applied to dramatic poetry (1874).

THE SOUND OF MUSIC



Use of a 3-Facet Rasch Model to Measure Small Audio Impairments in the Field of Audio Engineering

David Moulton and *Mark Moulton, Ph.D.*
Moulton Laboratories, Groton, MA, and San Jose, CA

Abstract

The subjective measurement of small audible differences in the audio engineering field has been hampered by experimental conflicts between applicability and reproducibility. The Rasch Model offers a powerful means of controlling the statistical analysis of experimental data in order to maximize reproducibility and applicability across listeners, audio material, and devices under test. The authors describe their testing of five perceptual audio coders for Lucent Technologies.

The Problem of Measuring Perception of Small Audible Impairments

Measurement of listener perception of small audible impairments caused by audio reproduction devices has been constrained by the combined but conflicting needs for (a) reproducible test results and (b) broadly applicable conclusions. Measurement techniques have sought to achieve reproducibility through rigorous test design and execution intended to minimize such sources of uncontrolled variance as listener training and expertise, the choice of program material, and the listening environment. For example, only expert listeners are used and the listening environment must meet exacting specifications. This poses a dilemma. The more rigorously controlled the testing environment, the less faithfully it reflects the listening conditions of the real world. Most listeners are not experts. Most rooms do not meet the specifications for properly controlled listening environments.

Test data are drawn from rating scales such as the Mean Opinion Scale (MOS)[1] and often incorporate an accuracy test in which the listener must pick out a reference signal from among a selection. Analysis of Variance (ANOVA) is used to interpret the results.[2] Data collection rigor is presumed to minimize random statistical variance and to reduce systematic biases. Techniques such as diff-grade analysis are used to diagnose listener inexpertise and to reduce rating scale floor and ceiling effects.[3] However, even under perfectly controlled test conditions, anomalies arise that compromise reproducibility and that ANOVA is not competent to remedy. We encountered several such instances in our study.

The conventional experimental approach is drawn from procedures traditionally used to control objective data from which the human element has been removed. As a consequence it rests on several assumptions that are hard to support. First, it assumes that all extraneous sources of variation can in fact be experimentally removed so that what is revealed are the perceptions themselves and not biases of the listeners, characteristics of the audio systems, or anomalies arising from particular cases. However, the physical and psychological complexity of the listening process appears to render this level of control impossible at the laboratory level. Even under the most controlled conditions, researchers have found replication to be extremely difficult.[4]

Second, it assumes that test subjects unequivocally perceive and can identify the small impairments under test, in other words that they are "experts." Researchers attempt to meet this condition through a process of pre-screening listeners and post-test removal of "non-experts" who fail to meet a guessing accuracy criterion. In reality, of course, listeners bring a continuum of expertise and perceptual acuity to such tests, and no listener is sufficiently expert to produce the kind of reliable measurements ultimately desired. There is also the problem of relating the reports of experts to the probable experience of non-experts. A hypothetical panel of "perfect" experts would lead one to conclude that even the best perceptual audio coding systems are "extremely annoying," leaving fully open the question of how such systems would be perceived by the rest of the world.

Third, there is an assumption that such perceptions can be reduced to a reliable, stable, and reproducible metric, that they are in fact measurable to the point where they may be quantified in a useful way for subsequent use in the design, manufacture, and application of audio systems.[5] It is well known that rating scale data do not possess these metric properties.[6] The relative spacing of the rating scale categories is highly variable and there are pronounced compression effects at the top and bottom of the scale, making it highly nonlinear. While use of diff-grades has made such difficulties more manageable, the fact remains that a rating scale is not a measuring stick.

Fourth, it is assumed that Analysis of Variance is suitable for this type of analysis. However, ANOVA specifies: 1) linear, interval scales; 2) representative samples; and 3) an absence of interaction effects if the intent is to measure main effects. None of these specifications is met in this type of data. The scales are nonlinear. The expert listeners represent no population but their own. Interaction effects abound, and while ANOVA can be used to document their presence, it can do little to prevent their perturbation of the main effects. As a consequence, results drawn from ANOVA do not reproduce well when the selection of programs or listeners is changed.

The Listening Format and Devices Under Test

The devices under test were five high-performance Perceptual Audio Coders known as "codecs." Perceptual Audio Coders are complex encoding algorithms used to remove data from a digital audio signal for ease and speed of electronic transmission. They are "perceptual" in the sense that they take advantage of the physical and psychological mechanics of hearing perception to identify means of removing information from a sound signal in such a way that the brain does not detect the loss. An enormous amount of audio data can be removed before the brain senses anything missing, but eventually as data is removed the brain hears "glitches" in the audio signal. It was the purpose of these tests to measure the audibility of such "glitches" for a specific codec that Lucent Technologies hopes to use in the field of digital radio broadcasting. (Radio broadcasting currently uses "analog" signals which lack the flexibility and wide applicability of digital signals.)

The authors measured the five codecs using a panel of thirty listeners with a wide range of experience (we deliberately included nonexperts) and other demographic characteristics, and ten audio examples drawn from commercial and test recordings. All testing was double-blind and done in small groups over a two-month period, using headphones. The goal of the test was to determine the relative impairment each codec contributed to reference recordings for a range of listeners listening to a range of conventional recordings.

The test consisted of fifty examples, following a training session and three warm-up examples. Each example consisted of a sequence of recordings identified as "Reference", "A," "B," "again, Reference," "A," "B." In each case, the identified reference was one of the Reference recordings, while A or B was the codec-processed copy under test and the remaining of A or B was the reference again (the so-called "hidden reference"). The listeners were asked to score both A and B according to the given criteria, and to identify which of A or B was the hidden reference.

There were two tasks: 1) rating each codec on the 5-point Mean Opinion Scale; 2) picking out the hidden reference. In a conventional diff-grade analysis, the two tasks would be combined into one set of "ratings." The listener would automatically assign a "5" to his *guess* of the hidden reference. The diff-grade would then be the difference between the rating given the *actual* hidden reference and the rating given the encoded signal. These diff-grades would be used to screen out

non-experts. For the Lucent test, listeners were not forced to assign a "5" to one of the choices since diff-grades were not used. Instead, we simply performed two distinct but parallel analyses, the first using the MOS ratings to measure codec transparency, the second using frequency of correct identifications of the hidden reference.

The Mean Opinion Scale was presented as follows:

5 = I cannot hear a difference between the reference and the processed recordings.
4 = I hear a perceptible but not annoying difference between the reference and the processed recordings.
3 = I hear a slightly annoying difference between the reference and the processed recordings.
2 = hear a distinctly annoying difference between the reference and the processed recordings.
1 = I hear an extremely annoying difference between the reference and the processed recordings.

Following the test session, listeners were asked to complete an exit questionnaire. To the question, "Were the PACs in general hard to distinguish from the reference signal?" 27 (90%) answered yes, and 3 (10%) answered no.

Theoretical Justifications for Using a 3-Facet Rasch Model

To analyze the ratings we employed a 3-Facet Rasch Model.[7] Each datum was conceived to be the conjoint effect of the "transparency" of the Codec under test, the "severity" of the Listener, and the "intolerance" of the audio sample or Program to Codec artifacts. The corresponding expression, including an F term to take into account transitions between adjacent categories, was:

$$P\{x_{CLMF} \geq "k" | C_n, L_i, M_j, F_k\} = \frac{e^{C_n - L_i - M_j - F_k}}{1 + e^{C_n - L_i - M_j - F}}$$

where x_{CLMF} = the rating value assigned a Codec
k = a rating scale category
C_n = transparency of Codec n in logits
L_j = severity of Listener i in logits
M_j = intolerance of Program j in logits
F_k = difficulty of the step up from category k-1 to k

Equation 1

In other words, the probability that a given response x will be greater than or equal the k'th rating scale category given Codec C, Listener L, Program M, and step difficulty F of reaching k from k-1, is a function of the logit measures of C, L, M, and F.

The Logit Scale

It will be recalled that conventional subjective testing assumes a stable, linear metric, a condition that is not met by the MOS scale. First, rating scales that have a clear "floor" and "ceiling" such as the MOS scale, whose ratings must fall between "1" and "5," suffer compression effects at the end of the scale. Such effects are ameliorated by using only the middle categories of the scale (not practicable with high-performance codecs) and by using diff-grades, where each rating is replaced by the difference between the rating given the Codec under test and that given a Reference signal. (Diff-grades cleverly smooth out the ceiling effect by introducing the possibility of extra categories at the top of the scale arising from incorrect identifications of the Reference signal, which are then discarded as unreliable, thus locating the set of "reliable" responses towards the center of the diff-grade scale.) The second reason why the MOS metric is not preferred is that, compression effects aside, the length of each rating scale unit depends on the relative wording of adjacent category descriptions, which is highly variable, creating a ruler without consistent units, for which no "centimeter" matches any other.

Rasch measures meet the demand for a stable, linear scale by replacing the MOS rating metric with the logit scale which measures distance in terms of linearized probabilities—the log of the probability of scoring above a specified category divided by the probability of scoring below it. The logit scale suffers no floor or ceiling compression effects as it has no upper or lower limit, and each logit is the same "size" as every other. It can also be readily interpreted as the probability of a particular codec scoring at or above a specified rating when confronted with a listener of a given severity and a program of a given intolerance. Thus, it now appears possible for the audio field to measure perceptual audio coder transparency in a metric as useful and definable as the decibel (which measures loudness on a similarly logarithmic scale) and the other physically defined variables that characterize sound.

Unidimensionality

An important feature of the Rasch Model is that it requires unidimensionality of test items as a condition of fit. Yet all data sets, including the one analyzed here, are multidimensional to some degree, no matter how careful the researchers. What, then, of the Model's applicability? So long as there is a single *dominant* dimension, such as Codec Transparency, the Model is applicable. Extra dimensions manifest as misfit and are purged from the data set accordingly. Thus, unidimensionality is an ideal which the Model tests for and makes it possible to approach. It is not a precondition of successful analysis.

In comparison with the educational and psychological data to which the model is routinely applied, the audio data

set analyzed here was found to be exceptionally unidimensional.

Editing the Data Set to Maximize Reproducibility

Measure reproducibility is the biggest obstacle faced by the audio industry in trying to determine the quality of audio devices. Codecs seem to perform differently in different testing situations no matter how rigorous the testing environment. Part of the problem has been an inability to specify what is meant by reproducibility and to edit data sets to maximize it. The very idea of "editing" a data set sounds heretical from a statistical point of view, and rightly so, not just because ANOVA and other statistical techniques require complete data but because editing compromises the random nature of the sample and thus its representativeness of a larger population. A sample-independent model like Rasch, however, makes no assumptions regarding randomness or representativeness, and it does not require complete data. Indeed, the model is in some senses not a statistical method at all. It merely specifies how data must behave in order to lead to reproducible measures. The data must behave as if attributable to objects that occupy a single position on a single unidimensional scale.

Rasch generates two types of numbers. The first are the logit measures and associated output which correspond to each Codec, Listener, and Program. The second are the expected values expressed in the rating scale metric which are computed for each cell of the data matrix from the logit measures and compared with the corresponding actual data values. It is the summation of their residuals across a set of cells which becomes the basis of the fit statistics associated with each Codec, Listener, and Program.

Suppose, then, we see Codec A misfitting significantly. In conventional Analysis of Variance not much can be done. We can remove Codec A from the analysis, but that leaves us nowhere. We can treat Codec A's overall measure as a Main Effect, then look for the interactions with particular Programs and Listeners that might be causing the misfit and report these as Interaction Effects. But the more pronounced the interaction effects (or biases, for that is what they are), the less trustworthy are the main effects. Could we, then, recompute the main effects after removing the interaction effects? Unfortunately, no. Since ANOVA depends strongly on complete data, on having no missing cells, there is no way to remove the data causing the interaction effects without significantly compromising the interpretability of the results. In short, while ANOVA can offer a diagnosis, it does not supply a cure.

Since it separately models each cell in the data matrix, Rasch does not require complete data. That means the misfitting cells causing interaction effects can be suspended from future analyses (treated as "missing") without compromising the interpretability of the results. The methodology thus implies an iterative process of suspending misfitting data from the analysis (filing it away for diagnostic purposes), recomputing the measures and expected values, identifying and removing the new crop of misfitting cells, recomputing the measures, and so forth. The process is concluded when there are no longer significant misfits, or in other words *when the main effects have been completely purged of interaction effects*.

A full treatment of the relationship of Rasch to ANOVA has yet to be attempted, particularly with respect to the Main Effects/Interaction Effects contrast. I think such would prove enormously valuable to the many fields which, like audio engineering, rely almost exclusively on ANOVA and related methodologies to interpret results. Unable to subtract interaction effects mathematically, researchers must labor to remove them physically from the experiment, often futilely and at great cost.

Results of the Analysis

We performed two parallel and independent analyses, the first to measure codec transparency by a rating scale analysis of the MOS ratings, the second to measure transparency in terms of listener inaccuracy. Since the two forms of analysis are independent methods of looking at the same construct (codec transparency), we felt that a comparison of the two sets of results would act as a cross-check on their reproducibility. Strong agreement would suggest a high likelihood of reproducibility and was in fact found. The correlation between measures derived from the MOS ratings and those derived from the listener's ability to pick the Reference signal in each A/B pair, was 0.99, a pure straight-line relationship, regardless of the fact that the two data sets are substantially independent of each other.

This paper focuses on just the MOS rating scale analysis.

The Rating Scale Analysis

Table 1 gives the MOS generated logit measures for the Codecs and two Reference signals (where the signal suffered no audio coding) after the significant biases were removed (i.e., bias z-score > +2.0 or <-2.0). (The biases themselves and their probable effects on Codec perception will be discussed shortly.) The relative positions of the logit measures in **Table 1** should be very close to those that would be calculated using a different panel of listeners and a different set of audio samples, provided the biases are removed from these as well. The Separation statistic for the codec measures is 8.56, indicating that the codecs have been reliably distinguished by the listening panel. The fact that they are significantly different from the

"Ref" measures (which are computed from ratings given the hidden reference) tells us that the listening panel as a whole was able to reliably detect even the best codecs.

Table 1: Codec measures, biases removed

Codec	Logit Transparency	Model S.E.	Fair Avrge	Misfit
Ref1	3.24	0.17	4.80	1.20
Ref2	3.01	0.15	4.80	1.00
Codec1	2.09	0.11	4.50	1.00
Codec4	2.07	0.13	4.50	1.00
Codec2	1.89	0.12	4.40	1.00
Codec3	1.11	0.10	4.10	0.90
Codec5	-0.28	0.09	3.00	0.90
Mean	1.87	0.12	4.30	1.00
S.D.	1.10	0.03	0.60	0.10

Looking across the top of **Table 1** at the column headings:

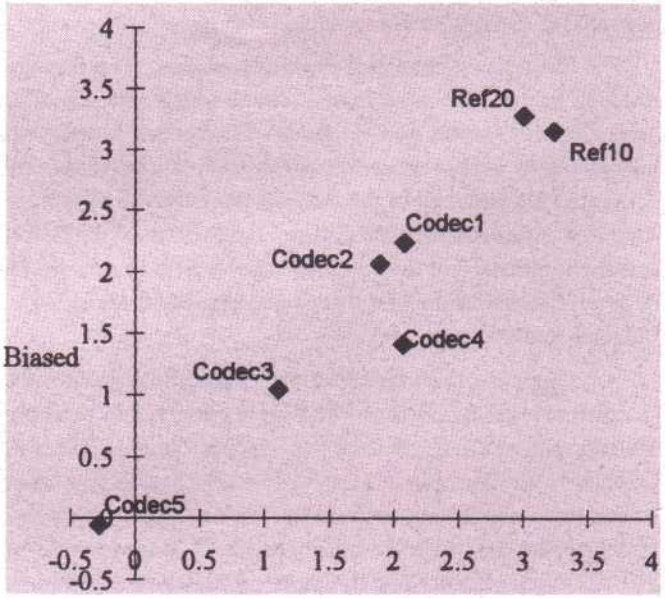
- The **"Codec"** column gives the labels for the codecs analyzed. "Codec4" refers to what we eventually learned was Lucent's PAC at a 96 kb/sec, the audio coder that Lucent plans to use for digital broadcasting. Notice that it performed almost as well as Codec 1 which uses 128 kb/sec, quite a lot more audio information. "Ref1" and "Ref2" are based on the ratings that were given unknowingly to the reference signals when they were compared to Codecs 1 and 2. (Listener comparison of the reference distracters with Codecs 3, 4, and 5 was too easy, artificially inflating their mean MOS scores and creating significant misfit, justifying their exclusion from the analysis.)
- The **"Logit Transparency"** column gives the codec transparency measures on a logit or log-odds unit scale (higher means "more transparent") from which probabilities can be computed using Equation 1. Since the zero point of the scale is arbitrarily set at the mean "severity" level of the Listeners and the mean "intolerance" level of the Programs, the codec probability of scoring "4" (audible but not annoying) or better for the average Listener and Program is easily calculated as: exp(codec measure)/(1 + exp(codec measure)).
- **"Model S.E."** is the standard error in logits of each codec measure, computed assuming the data "fit" the model, an assumption supported by the Misfit column shown next.
- **"Fair Avrge"** is the average of the Rasch expected values for that codec, expressed in the rating scale metric.
- **"Misfit"** is the ratio of observed to expected noise in the estimate and is ideally 1.0. It is calculated as the mean of the squared residuals divided by the variance of the estimate.

On the basis of this table, it was found that Lucent's PAC at 96 kb/sec (Codec 4) was as transparent as Codec 1 which uses 128 kb/sec, a genuine feat of encoding. Applying **Equation 1**, we can say that the average listener (in our sample) listening to the average program will rate Codec 4 "perceptible but not annoying" or better 89% of the time.

Effect of Including Biases

Table 1 is the result of an iterative process of removing biases and interactions between the codecs, programs, and listeners and recalculating parameters. Its virtue is that subsequent analyses with different programs and listeners should result in very similar codec measures, so long as they undergo the same process of removing biases and interactions. However, it does not reveal the peculiarities of this particular test administration. For that, Bias tables (not shown here) are used which show the precise size of the interactions between codecs, programs, and listeners. **Figure 1** graphs the codec measures with biases removed against the codec measures when they have not been removed. You will see that the two sets of estimates are quite similar, with one startling exception.

Figure 1: Biased vs. Unbiased Codec Measures



Unbiased (Rescaled to have same Mean, SD)

Notice that the measure for Codec 4 drops significantly, to third place, when biases are included. A perusal of the bias statistics reveals that approximately 75% of this drop is due to an interaction between Codec 4 and the Castanets audio sample. Reproducing the other samples, Codec 4 performed extremely well. Reproducing an audio sample featuring sharp, percussive castanets, it performed poorly, uncharacteristically so. This provided valuable information to Lucent Technologies, enabling it to identify and remove an error in the encoding algorithm which was causing the castanets interaction.

This raises an important question, of course. Which is the "correct" measure of Codec 4? The answer depends on the goal of the researcher. If the goal is to create reproducible measures, measures which are the same from one testing situation to another, the "correct" measure is the unbiased one—so long as all tests are subject to the same iterative process of removing biases and misfits. If the goal is to describe the effects of a particular testing situation, the biased measure is more reflective of what happened, although it is better to explore biases individually than through their effects on an average.

Listener Severity

The Rasch Model computes estimates for Listeners and Programs at the same time that it estimates Codec transparency. Because the Model makes no assumptions regarding the nature or distribution of the sample, there is no need for Listeners and Programs to be normally distributed along the variable. In fact, it can be seen that the Listener distribution is bimodal, dividing cleanly into "experts" and "non-experts." **Tables 2** and 3 provide Listener and Program measurements.

Table 2: Listener Severity

Listener	Severity	Model	Fair Avrge	Misfit
10/Lro2255	1.04	0.16	2.3	1
5/Gle2255	0.74	0.17	2.6	0.6
8/Mib2253	0.63	0.17	2.7	0.6
1/Eos3255	0.44	0.22	2.8	0.6
2/The3154	0.36	0.18	2.9	0.9
20/Eys3244	0.31	0.18	3	1.2
21/Gla2154	0.25	0.22	3	1.3
22/Nar3144	0.2	0.18	3.1	1.3
24/Har3253	0.2	0.18	3.1	1
28/Dri4243	0.15	0.23	3.1	0.8
30/Moi4255	0.13	0.18	3.1	1
12/Gra1111	0.12	0.19	3.1	1.3
23/Shi2254	0.03	0.19	3.2	1.3
3/Ace4111	0	0.24	3.3	0.8
27/Tin2251	-0.18	0.24	3.4	1.2
25/Urr2113	-0.58	0.22	3.7	0.7
9/Utt2113	-0.63	0.22	3.8	0.6
15/Hra1212	-0.63	0.23	3.8	0.9
26/Gul2133	-0.64	0.23	3.8	0.6
29/Cre2244	-0.65	0.28	3.8	0.9
7/Cou4131	-1.28	0.28	4.2	1.5
Mean	0	0.21	3.2	0.9
S.D.	0.55	0.03	0.5	0.3

Table 2's first column lists each listener with a name abbreviation and background code. The first two digits of the code give their gender (1 = Female) and age (5 = ">50"). The last two digits indicate audio and musical experience where "5" means "extensive training and experience." Notice that the experts cluster toward the top, at the severe end of the scale, the non-experts toward the bottom. Experts are better able to discern audio artifacts, making them more likely to use the lower categories of the scale.

Note also that there are only 21 Listeners listed, though data was gathered for 30. The remaining nine were suspended from the analysis due to high misfit, indicative of internally contradictory response strings. The fact that many of the remaining listeners are non-experts, as evidenced both by their background and their lack of severity, indicates that it is possible to generate reliable measures using non-expert listeners. Because these listener measures are on the same logit scale as the codecs, and because they have been linked to the general population through background demographic information, it becomes possible to make predictions regarding the perception of codecs for the larger population for which they are intended. For instance, taking the average severity measure of those with combined expertise scores of less than 6 as derived from a brief entrance questionnaire, and putting it through **Equation 1**, we find that non-experts (those with little or no musical and audio experience and training) have a 93% chance of finding Codec 4 to be "Perceptible but not annoying" or better. In fact, we can compute the probability that any potential listener will find Codec 4 to be annoying without administering a listening test at all. We need only ask a few questions about musical and audio background and apply a regression equation to predict listener severity, from which probabilities can be computed, a procedure described in another paper.[8]

We can therefore claim that the need for measures having relevance to the larger listening population has been met using only a small, unrepresentative panel of listeners.

Program Intolerance

Finally, let us consider the measurement of Program Intolerance.

Table 3: Program Intolerance to Codec Artifacts

Program	Intolerance	Model	Fair	Misfit
Malespeaking52	0.94	0.14	2.4	1.1
Ethridge1152	0.23	0.13	3	1.1
US3/3342	0.19	0.13	3.1	0.9
B52s2343	0.18	0.14	3.1	0.8
Fagen2233	0.02	0.13	3.2	0.9
Chicago4334	0.01	0.13	3.2	0.9
Sweet Honey2123	-0.08	0.14	3.3	0.9
Castenets4513	-0.15	0.15	3.4	1
Folger2115	-0.52	0.15	3.7	1
Berlioz3115	-0.82	0.16	3.9	0.9
Mean	0	0.14	3.2	0.9
S.D.	0.44	0.01	0.4	0.1

The program column of **Table 3** contains the audio sample used and a code of acoustical characteristics—Dynamic Range, Crest Factor, Distortion, and Reverberence. Interestingly, as the intolerance of the programs to codec artifacts moves up the logit scale, the Reverberence rating decreases from "5" to "1." This suggests that reverberance covers up codec artifacts and can in fact be used to predict program intolerance, just as audio and musical experience can be used to predict listener severity. This was a finding not anticipated by the test administrator. Thus, it is theoretically possible to compute the probability that a given codec will be annoying just by measuring the reverberance of the audio signal electronically.

Observe that the Castanets misfit is a perfect 1.0. This is because its interactions with Codec 4 have been removed. Originally the Castanets misfit was in excess of 1.6.

Conclusions

The Rasch Model shows promise as an inexpensive means of supporting and enforcing the experimental control of audio experiments by statistical means in order to generate reproducible measures. Indeed, in some respects it offers a level of control that extends beyond what could be achieved by ideal experimental conditions, as when it identifies biases and extraneous effects originating from the actual codecs under test. An example of this is the Castanets bias against Codec 4. Since the bias arose from a programming defect within the codec, no amount of experimental control could have prevented it. Without the measurement control imposed by the model, Codec 4's performance would have been doomed to vacillate along the Transparency scale depending solely on the accident of whether or not the Castanets program happened to be present among the sample of programs used in the test.

Are these Rasch measures in fact reproducible? The answer depends on future research, testing the same codecs at a different site using different listeners and programs. There are reasonable grounds for hope. First, we have a well-documented theory supported by extensive educational and psychometric research which finds that such measures will reproduce when a sufficiently diverse set of data have been found to define a coherent variable, i.e., when the data fit the measurement criteria of the model. Second, the reliability statistic for the codec measures, corresponding to a signal to noise ratio of 8.56, is 0.99. Third, a parallel analysis based not on how listeners reported perceiving the codecs, but on their actual success rates in identifying the hidden reference, generates codec measures which are statistically identical ($r = 0.99$) with those generated using the MOS audibility scale, again suggesting reproducibility. We feel that if such preliminary indications are borne out over time, the Rasch Model will prove a useful and cost-saving addition to audio testing methodologies.

Acknowledgments

We wish to acknowledge Søren Beck for his assistance in interpreting the ITU-R testing recommendations during our original research, Benjamin Wright for his advice on research design and use of the Rasch Model, Deepen Sinha for involving us in his codec development work, and Lucent Technologies for their strong ongoing support.

For more information, contact:
Mark Moulton
319-A Page Street
San Jose, CA 95126
(408) 279-1953
E-mail: 73014.340@compuserve.com

1The Mean Opinion Scale is usually a 5-point rating scale (5 = No perceptible difference, 4 = Perceptible but not annoying, 3 = Slightly annoying, 2 = Distinctly annoying, 1 = Extremely annoying). Its use is common in the audio industry.

2ITU-R Recommendation BS.1116. "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," 1996, Section 2. (The ITU is an international board used by the audio engineering field to set standards).

3In the accepted "Reference Signal /Test Signal A /Test Signal B" testing format, one of the two test signals is the reference signal replayed. The listener is supposed to guess which one, then assign it a "5" as having "No audible difference." He then rates the other signal on the 5-point scale. The "diff-grade" is the difference between the rating assigned the hidden reference and that assigned to the other test signal (Diff-grade = Hidden Reference rating - Other Test Signal rating). This is the metric recommended by the ITU. For more on diff-grades, see Section 3.0.

4Thomas Sporer, "Evaluating Small Impairments with the Mean Opinion Scale—Reliable or Just a Guess?" AES Preprint, November 1996.

5ITU-R Reccomendation BS.1116, Section 3.2.

6Benjamin Wright & Geoffrey Masters, *Rating Scale Analysis* (Chicago: MESA Press, 1982).

7 John M. Linacre, *Many-Facet Rasch Measurement* (Chicago: MESA Press, 1994), pp. 1-21.

8We describe just such an analysis performed using these data. David and Mark Moulton, "Codec 'Transparency,' Listener 'Severity,' Program 'Intolerance': Suggestive Relationships between Rasch Measures and Some Background Variables." Audio Engineering Society Preprint, 106th AES Convention, September 1998, San Francisco.

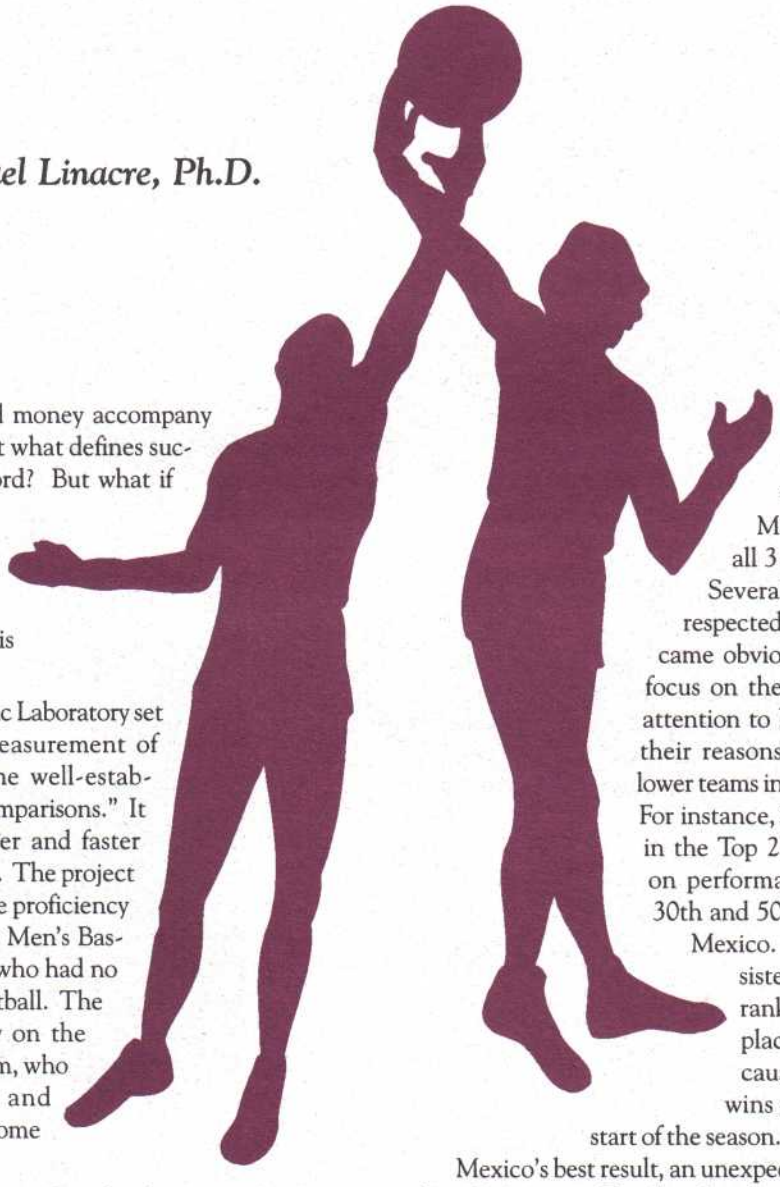
David Moulton:
Audio Engineer, Educator, and Author. Owner of Digital Media Services, a multitrack and surround post-production facility. Principal in Sausalito Audio Works, a firm licensing high-performance loudspeaker technology. Author of "Golden Ears Audio Ear Training" and "Total Recording" (1998 release). Degrees from Bard College and Juilliard School of Music.

Mark Moulton, Ph.D.:
Psychometrician. Specializes in statistical measurement and prediction using Rasch analytic techniques in a variety of fields including psychology, audio, and economic forecasting. Author of "n-Dimensional Replacement: Implications of a Rasch Geometry." Degrees from St. John's College and University of Chicago.

What Are The Odds?

Measuring College Basketball

John Michael Linacre, Ph.D.



Pride, prestige, and money accompany a successful sports team. But what defines success? A good won-loss record? But what if only weak opponents are played? Experts' opinions? But what if they fail to notice you? A simple, fair, objective measurement system is needed.

MESA Psychometric Laboratory set out to demonstrate the measurement of team performance using the well-established method of "paired comparisons." It turned out to be even easier and faster than was initially envisioned. The project involved measurement of the proficiency of 1998-9 NCAA Division I Men's Basketball Teams by an analyst who had no experience in College Basketball. The measures were based solely on the won-loss records of each team, who their opponents were, and whether the games were at home or on the road.

The results were astounding for their immediacy, simplicity, and face validity! Basketball games are played almost every day from November to March. MESA published team measures, updated daily, for the entire basketball season. MESA's top 20 teams were generally the same as those of the Associated Press weekly poll of 70 basketball experts — though with minor differences in ordering. AP, however, only lists the Top 25 teams. MESA measured and ranked all 315 Division I teams.

Several idiosyncracies in the widely respected and reported AP system became obvious. The basketball experts focus on the top teams. They pay less attention to lower teams. Consequently their reasons for choosing some of the lower teams in the Top 25 are idiosyncratic. For instance, Syracuse was always ranked in the Top 25, yet MESA ranked them, on performance, consistently between 30th and 50th. Even stranger was New Mexico. AP ranked this team consistently in the Top 25, MESA ranked them around 75th. AP placed them in the Top 25 because they had a run of home wins against weak teams at the start of the season. This was reinforced by New Mexico's best result, an unexpected home win against Number 13, Arizona. The fact that they were beaten on the road by Number 242, Hawaii, seems to have been ignored. It seemed that AP experts were reluctant to drop teams from the top 25 or introduce new teams. It was not until Florida was the 10th best team (according to MESA), that the AP experts voted it into the top 25.

The real difference between AP and MESA is in terms of prediction. MESA provides every team a measure of proficiency on a linear scale. The measurement model for a win by Home Team, H, over Guest Team, G, is the Rasch paired comparison model:

log (Probability (Win by H) / Probability (Loss by H)) = H's Proficiency + Home Court Advantage - G's Proficiency

Each day, MESA computed a measure for each of the 315 teams and the size of the home court advantage. This made it possible to predict the outcome of each night's basketball games. 65% of basketball games were won by the home team. The home court advantage corresponded to .8 logits. The range of measures of Div. I teams covered 10 logits. Since MESA computed a measure for every team, the victor of a game was predicted to be the larger of (Home team strength + Home court advantage) and (Guest team strength). For games played on neutral courts, there was no home court advantage. MESA successfully predicted, in advance, 72% of game outcomes, i.e., about 3 out of 4 games. This performance is on a par with professional tipsters, but they only predict selected games, not the entire schedule. On 3 days, MESA correctly predicted all game results. The worst showing was one day when only 40% of outcomes were predicted correctly.

This same simple technique of paired comparisons has been applied in many contexts. Here are the steps that were followed for NCAA Basketball:

1. Download a list of teams. The definitive list of Division I teams was found on the NCAA web site a few weeks into the season. Initially a list was built up from reported results. Since some teams were found to have variants to their names, a synonym list of team names was constructed. New teams were also added to the list as Division I teams played other schools. These were added to keep the won-loss records correct, but had little influence on measures.

2. Since there are many teams which maintain perfect records for a few games at the start of the season, a win against a notional very bad team and a loss against a notional very good team were imputed for each actual team. Pre-season rankings were also incorporated, but these were found to become uninformative after each team played only a few games.

3. Download accurate results daily. Many sources provided results for the AP Top 25. Yahoo alone provided scores for all games and indicated home team. Mistakes and omissions, however, occurred. Checking the won-loss records of top teams against their own websites prevented conspicuous blunders. Since it was not always obvious who were home teams at invitational and tournament events, some detective work was required. Measures, however, proved to be robust against occasional reporting and data-entry errors.

4. Add current results to the database of cumulative results and estimate measures. A series of short BASIC programs edited, checked, and formatted the downloaded results into a data file suitable for analysis by the Facets program. Analysis was performed and another BASIC program reformatted the Facets output into HTML web pages which could be immediately uploaded onto the MESA web site. Turn-around time was less than an hour. In situations in which there is no home court advantage nor order effect in the paired comparisons, then it would be easier to use more conventional Rasch software, such as WINSTEPS. 5. Make predictions. The daily schedule of games could be downloaded and results predicted. Since all teams and the home court advantage were measured, making predictions was a "piece-of-cake."

MESA encourages others to apply these techniques to sports competitions, consumer preferences, value comparisons, or other situations in which outcomes are based on paired contrasts. Mike Linacre, MESA Psychometric Laboratory University of Chicago MESA@uchicago.edu

John Michael (Mike) Linacre, Ph.D.

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

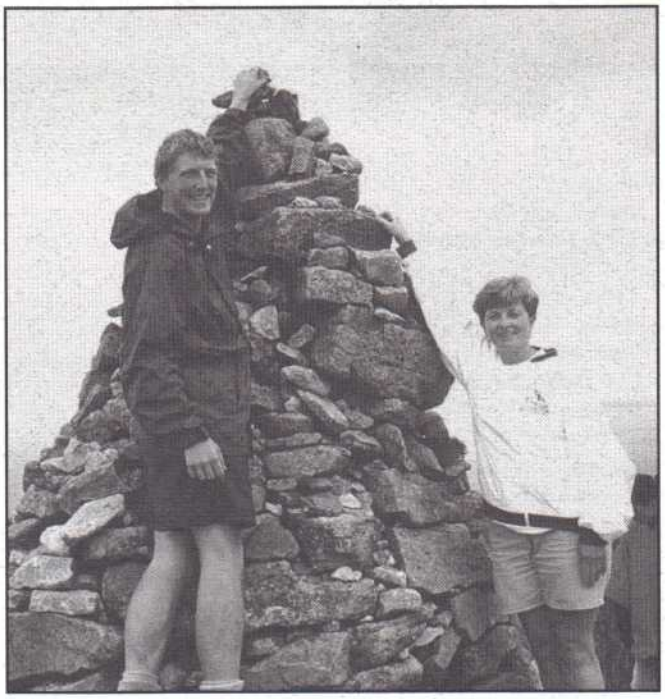
Measuring Mountains

Ryan Bowles
Computer Adapetive Technologies

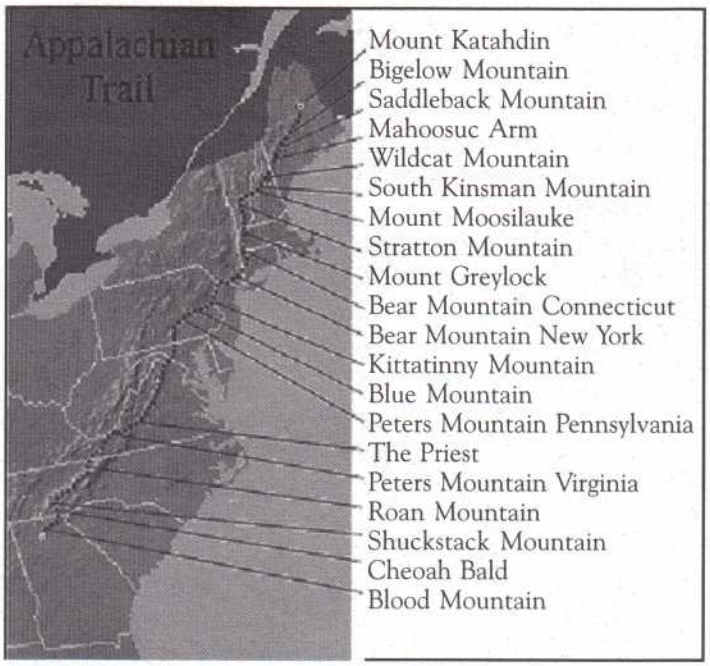
Ryan Bowles at end of trail.

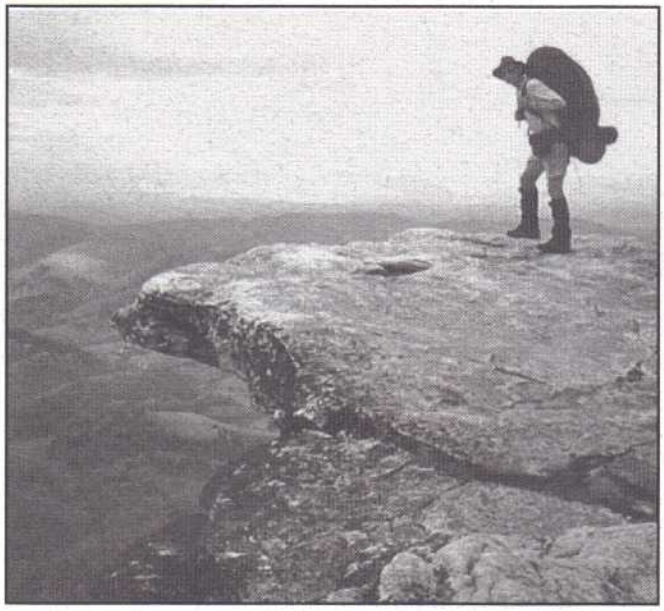
I hiked the entire Appalachian Trail in the spring and summer of 1997. I started in Georgia on Springer Mountain on March 16 and finished in Maine on top of Mount Katahdin on August 6. The Appalachian Trail is a footpath, 2160 miles long, running through woods and fields, along ridgetops and streams, through 14 states. The Trail goes up and down over many mountains, ranging in elevation from 142 feet above sea level on the Hudson River in New York, to 6643 feet on Clingman's Dome on the border of North Carolina and Tennessee. Every year over 2000 people attempt to hike the entire Trail in one season, a feat known as a thru-hike. Only about 10 percent are successful. Many different kinds of people try to thru-hike the Trail: old and young, fit and overweight, factory workers and company executives. Within a few weeks, though, these differences have disappeared and everyone is in nearly perfect shape. Even so, thru-hikers' perceptions about how difficult it is to climb a mountain differ because of such variables as weather, tiredness, and pack weight. I was curious about the difficulty of the mountains along the Appalachian Trail, but had no way to remove these idiosyncrasies, until I came across Rasch analysis.

I asked 60 people who had hiked the entire Appalachian Trail to rate the difficulty of twenty mountains scattered



Ryan and another hiker





Author at a high point in his journey

along the length of the Trail. All had hiked the Trail from south to north within the last ten years. For each mountain, they said whether they thought the mountain was easy, medium, hard, very hard, or extremely hard. This survey was administered over e-mail, and responding was voluntary. Although my sample was not random, results from Rasch analysis are sample independent. I ran an analysis using the computer program Bigsteps. The following table lists the twenty mountains in order of difficulty according to the Rasch analysis, along with some information about each mountain.

The column labeled "Diff. Meas. 1 to 10" shows the measure of difficulty from the Rasch analysis, on a scale of 1 to 10, with 1 being the difficulty of the easiest mountain on the list, Kittatinny Mountain, and 10 the hardest, Mount Katahdin. It is important to note that this is an equal interval scale. That is, we can consider the difficulty measure as a scale where the difference in difficulty between Mount Katahdin and Mahoosuc Arm, .86 difficulty units, is the same as the difference in difficulty between Blood Mountain and Peters Mountain (Pennsylvania).

The Infit and Outfit columns show measures of fit, or how much disagreement there is among the responses of the hikers. The expected value of both statistics is 1. We can see that three mountains are quite a bit different from 1 in either statistic: Bigelow Mountain, Blood Mountain,

Table 1:
Mountains along the Appalachian Trail in order of Difficulty

Mountains	State	Elevation in feet	Elevation Gain in feet	Distance of Ascent in miles	Diff. Meas. 1 to 10	Infit	Outfit
Mount Katahdin	Maine	5260	4100	5.2	10.00	1.65	1.57
Wildcat Mountain	New Hampshire	4400	2400	4	9.70	0.85	0.89
South Kinsman Mountain	New Hampshire	4300	2000	2.5	9.68	1.05	1.00
Mahoosuc Arm	Maine	3800	1600	1.4	9.14	1.04	1.17
Cheoah Bald	North Carolina	5500	4000	8.2	7.76	0.93	0.99
Mount Moosilauke	New Hampshire	4600	3500	5.5	7.69	1.16	1.04
Roan Mountain	Tennessee	6300	2200	2.7	7.63	0.96	1.22
Bigelow Mountain	Maine	4200	2950	6.8	6.82	0.49	0.48
Saddleback Mountain	Maine	4100	2500	5.6	6.73	0.89	0.82
Shuckstack Mountain	North Carolina	3800	2000	3.4	6.36	1.07	1.12
Stratton Mountain	Vermont	3900	1700	3.8	5.91	0.70	0.76
Blue Mountain	Pennsylvania	1500	1000	1.2	5.70	0.97	0.96
The Priest	Virginia	4100	800	1.2	5.32	1.23	1.22
Peters Mountain	Virginia	3300	1600	2.6	4.94	1.05	1.11
Mount Greylock	Massachusetts	3500	2500	7.5	4.88	0.81	0.85
Blood Mountain	Georgia	4500	650	1.3	4.81	1.42	1.42
Peters Mountain	Pennsylvania	1300	900	2.1	3.95	0.80	0.82
Bear Mountain	New York	1250	600	1.6	2.75	0.84	0.87
Bear Mountain	Connecticut	2300	1600	5.5	2.67	0.97	1.01
Kittatinny Mountain	New Jersey	1200	900	2.7	1.00	1.02	0.83

and Mount Katahdin. Blood Mountain has high fit statistics, indicating more disagreement than expected. Blood Mountain is the first major mountain on the Trail, about 28 miles from the start. Thru-hikers have not yet gotten into nearly perfect shape, and there is great variety in ability, explaining the disagreement. The infit of .49 and outfit of .48 for Bigelow Mountain indicates that hikers agreed more than expected on the difficulty of the mountain. I am at a loss to explain this. Mount Katahdin had even higher disagreement than Blood Mountain, with infit of 1.65 and outfit of 1.57. There are three parts to the explanation of the variability in responses. First, Mount Katahdin gets the worst weather of any mountain on this list. Second, when you get to the top of the mountain, you must turn around and go back the way you came, so some people do not carry packs. Third, since the five-mile climb up to the summit is the last five miles of a 2160 mile thru-hike, the emotional aspect varies. Some hikers are elated to be done finally, while others are depressed at losing this great adventure, and emotions affect how easy anything seems. These three sources of differences in experience explain why hikers have the most disagreement about the difficulty of Mount Katahdin.

When I got to the top of Mount Katahdin, I was exhausted. I thought it was the toughest climb on the entire Trail. At the top of Mount Katahdin, I reached the sign marking the end of the Trail. After nearly five months of hiking every day, I was done, and I was ecstatic. I had seen the tops of hundreds of mountains. I had met thousands of people, some out on the Trail for a day, some out for months. I had encountered wildlife, including two rattlesnakes, a porcupine, and two bears. I had observed a large portion of the United States, in such close detail few have seen. I had taken about five million steps to get to the top of what Rasch analysis has shown to be the hardest mountain on the Appalachian Trail. No wonder I was tired!

Ryan Bowles is originally from New Freedom, Pennsylvania, and presently lives in Chicago. He is a Program Associate at Computer Adaptive Technologies in Evanston, IL, and is pursuing a Ph.D. in Economics at the University of Chicago. In his spare time, Ryan enjoys reading great literature, visiting strange places, and hiking long trails.

Here's my stats song:
(sung to the tune of On the Road Again)
Doing Rasch again
I won't do traditional stats again
Not having linearity is a sin
I just can't wait to calibrate again.

Assessment:

Coming Of Age

Sherwyn P. Morreale, Ph.D.
Associate Director, National Communication Association

Philip A. Backlund, Ph.D.
Dean, Central Washington University

In the 1970s, the U.S. educational system begot the as-sessment movement. Many thought — and perhaps hoped — it would be a passing fancy, an educational fad that would fade away, if effectively ignored. But during the 1980s, assessment grew to adolescence and, like any troubled teenager, it had its supporters and detractors. More embraced by legislators and academic administrators, some faculty depreciated the assessment process as unnecessary and time-consuming, an inappropriate expectation of overburdened instructors and academic departments. But on to the 1990s! Assessment gurus emerged on the national and local scene and within disciplines, spawning conferences, workshops, and a cottage industry of consultants. Assessment developed and matured into a vigorous young adult, to such an extent that it is institutionalized at virtually every level of education. State legislatures mandate it; accrediting bodies require it; professional educational associations support it and create tools for its implementation; and teachers have begun to think of it is as good idea for themselves and their students.

If assessment has come a long way — and become an integral part of the educational endeavor — that raises a question as to whether it's a good thing or a bad thing. This article — written by two would-be gurus of assessment in the communication field — argues that assessment is a good thing by describing its benefits and providing a snapshot of what an effective assessment program might look like.

But first, we begin by clarifying some terms and processes for the benefit of the novice reader and as a point of departure for our discussion.

Clarifying Terms ... the "A" Words!

Assessment, accountability, and accreditation are activities conducted by or done to the academy that are related to one another and to the evaluation of the process, impact, and outcomes of education.

* Assessment is a process by which faculty and administrators evaluate the worth of one of their main activities: educating students. More specifically, assessment is a program of planned activities that includes tools and measurement devices which, when applied, evaluate student learning.

This definition provides a clue to why assessment has become institutionalized. It is a process by which instructors, departments, and educational institutions find out whether they are accomplishing what they intend in the classroom and in their educational programs. In an era when those processes are under scrutiny, assessment provides evidence of accomplishing one's pedagogical goals.

*Accountability is the broad process by which academic institutions, on behalf of the public, are held responsible by legislatures and other interested regional and local agencies. One part of accountability is being held responsible for providing evidence that students are learning what you claim they are learning, whether in a course, a department, or an entire school. Of course, schools are held accountable regarding factors other than student learning, such as fiscal responsibility, responsiveness to community needs, and the caliber of scholarship of their faculty.

Given this description, one can see that the results of assessing student learning inform the process of being held accountable. That is, if you prove through valid, reliable, and multiple assessment techniques that students are learning x, y, and z, and if you said that x, y, and z is what they should learn, then the results of assessing student learning are a useful part of the institution's accountability report.

*Accreditation is what happens for or to an academic institution or program — such as a teacher education program — if by being held accountable, the school proves it is carrying out its responsibilities efficaciously. Accreditation is granted by the U.S. Department of Education, and sanctioned by regional accrediting associations and discipline-specific organizations.

So assessing student learning can be viewed as supporting and informing accountability and accreditation. It has become an integral part of the educational fabric of our culture, in part, because of a desire for increased accountability in

education. Additionally, at any school and on any campus, assessment has other benefits that may not be immediately apparent to its detractors.

Benefits of Assessment

Legislatures, accrediting bodies, state boards of education, and internal reviewers all want to know if the education of students is having the desired effect. While the form of questions and requirements posed by these groups may vary, they seem to come down to six fundamental questions that are asked of faculty teaching courses, administrators chairing departments, and heads of academic institutions:

1. Who are you and why do you exist (Mission)? 2. What do you want to accomplish (Goals and Objectives)? 3. What procedures will you use to determine if the goals/objectives have been met (Assessment)? 4. What are the results of your assessment processes (Analysis)? 5. What changes will you make to your goals/objectives/outcomes/processes based on these results (Application of Results)? 6. What evidence do you have that this is a continuous cycle (Continuous improvement)?

As you can see from this list of questions, assessment is a circular process of educational program definition, review, and revision. It may be this circularity that interests agencies responsible for educational accountability. However, the processes also make good academic sense. Answering these questions provides a number of advantages for students, schools, and faculty and teachers.

Benefits to students.

When we answer the above questions, students reap the reward of a more dynamic and enhanced education. The end product, of which students are the beneficiaries, is constantly monitored, improved, and more responsive to their needs. Also, the results of assessing their learning can be shared with the students, so they can monitor and take pride in their own individual and collective achievements.

Benefits to schools.

When teachers have a clear idea of their school or institutional mission (and you may be surprised how many different concepts of "why we exist" are present on the average college campus), teachers are more able to act in concert with each other to meet that mission. When schools, departments, and teachers clearly describe their educational outcomes, then students, the public, and teachers themselves have a better sense of what students are to learn. This leads to more effectively designed educational programs and strategies, and therefore to overall improvement in the schools themselves.

Benefits to teachers and faculty.

The result of a dialogue about pedagogy and how to assess it is a better-informed and less competitive group of instructors, regardless of grade level. Not only is their end product improved, faculty who work together toward positive reform typically are more enthusiastic and committed and less defensive. The process of developing an assessment program together invigorates both academic content and the academic professional.

Characteristics of Good Assessment Programs.

Given that assessment is here to stay and has benefits for various stakeholders in education, an important question to ask is what it looks like when you are doing it well. Here is a top ten list of the characteristics of good assessment programs, derived from the literature of accreditation associations, academic campuses, and professional associations. Variations on this list have proven useful for developing assessment programs for courses, academic departments, and entire schools.

A Successful Assessment Program ....

1. Flows from an institution's mission, the educational purposes and department's mission and goals, and course-specific goals and student outcomes.
2. Emerges from a conceptual framework for student learning.
3. Is marked by faculty ownership, responsibility, and involvement.
4. Has institution-wide support.
5. Relies on the use of multiple methods and measures.
6. Supports equal access and equity, and honors diversity.
7. Provides feedback to students, teachers, and the institution.
8. Is cost-effective.
9. Leads to desirable and valuable change and improvement.
10. Includes a process for evaluating and assessing itself — the assessment program.

Sherwyn P. Morreale, Ph.D

Sherwyn P. Morreale (Ph.D., University of Denver, 1989) is an Associate Director of the National Communication Association. She is on leave from a faculty position with the University of Colorado at Colorado Springs. Her responsibilities with NCA include: staff liaison to the NCA governance boards, most particularly the Educational Policies Board, project officer on summer conferences, project director on communication education projects, and regular contributor to NCA's newsletter, Spectra. She also serves as the discipline's ambassador to interdisciplinary organizations, for example, the American Association for Higher Education, the Alliance for Curriculum Reform, the National Campaign for Freedom of Expression, the Department of Justice-Community Relations Service, among others. Morreale's research interests include all aspects of communication education, particularly public speaking, diversity, and communication competence and its assessment. She has authored or co-authored textbooks, journal articles, and book chapters

Is More Better?

Measuring the Effects of Full-Day Kindergarten

Donna Surges Tatum, Ph.D.

This is the dawning of the Age of Assessment. Legislatures, taxpayers, and parents are demanding accountability for the resources expended on our children's education. It is increasingly important for School Districts to conduct research. When schools implement new initiatives, they must be properly evaluated to ensure the best decisions concerning student learning. This research must use credible assessment tools which provide objective results to determine the efficacy of a program. We will examine an example of this proactive approach.

Current research in the neurosciences demonstrates the importance of early childhood education. The plasticity of children's brains in their first years of life mandates well-planned, well-executed pedagogy. Stimulating the neural pathways and building up strong networks in the brain has life-long implications. If schools provide early intervention and stimulus for students, it has great impact. Their readiness to learn is improved, increasing student progress.

Indianapolis Public Schools Superintendent Duncan N. P. (Pat) Pritchett, Jr. decided to test these ideas. In the fall of 1997 Indianapolis Public Schools (IPS) designed and implemented ten full-day pilot kindergartens to provide extended learning opportunities for general education students. Five of these classes are located in high schools, and five are in elementary schools. They are compared to five general education half-day kindergarten classes and thirteen Title I full-day classes.

Indianapolis Public Schools Superintendent Duncan N. P. (Pat) Pritchett, Jr.

In the summer of 1998, the District felt a sense of urgency in compiling data, as the Indiana State Legislature was considering a proposal to fund optional full-day kindergarten throughout the state. This report is prepared from data provided by Nancy E. Beatty, Title I Facilitator for the Indianapolis Public School District.

The analysis answers the following research questions:

1. Do students in full-day programs make greater gains in academic readiness and language than students in traditional half-day programs?
2. Do full-day kindergartners in classrooms located in high schools do as well as kindergartners participating in full-day programs located in elementary buildings?
3. Are full-day kindergarten programs as beneficial to typical students as it is to children who are attending full-day programs for compensatory purposes?
4. Are gains broad enough in scope and sufficient in magnitude to warrant the extended program?

Choose the Tools — Pre-Rasched for Best Fit

IPS used the Peabody Picture Vocabulary Test - Revised (PPVT-R) as one of the instruments to produce measures for their research. This is an excellent choice which demonstrates the strength of conclusions one can make with confidence when a tool is "Pre-Rasched."

The PPVT-R is an individually administered test of hearing vocabulary designed for persons 2-1/2 through 40 years of age who can see and hear reasonably well, and understand Standard English to some degree. In this sense, it is an achievement test, since it shows the extent of vocabulary acquisition. Though far from perfect, vocabulary is the best single index of school success.

The Peabody Picture Vocabulary Test was developed in 1959 by Lloyd M. and Leota M. Dunn. From a pool of 3885 words whose meanings could be clearly illustrated by black-and-white line drawings, the best 300 stimulus words and their decoys were chosen after careful and repeated field testing.

A revised edition of the PPVT was introduced in 1981. This version is significant because it uses the Rasch/Wright latent trait model to precisely calibrate the difficulty of each item. This information was used to construct the PPVT-R so it is equally sensitive at all ages up to adulthood.

Two parallel forms contain 5 training items, followed by 175 test items arranged in order of increasing difficulty. Each item has four simple, black-and-white illustrations arranged in a multiple-choice format. The subject's task is to select the picture which best illustrates the meaning of a stimulus word presented orally by the examiner.

Testing requires only 10 to 20 minutes, because the subject need answer only 35 to 45 items of suitable difficulty. Items that are far too easy or far too hard are not administered. Scoring, which is rapid and objective, is accomplished while the test is being administered.

The "Pre-Rasched" properties of the PPVT-R are important. The PPVT-R test items have been calibrated using the Rasch/Wright model. This makes the item calibrations independent of the student sample taking the test. The result is measurement of student ability in precise, linear, standardized units. This is important because direct comparisons can thus be made for student progress over time, and for particular groups.

Growth curves are constructed for hearing vocabulary. The normal development table converts raw scores to W-ability. The cumulative percentages for W-ability are then converted to normalized Z scores using tables based upon the normal probability curve. For each of the 25 age groups the Z scores were converted to unsmoothed normalized standard score equivalents.

Think in terms of a ruler. The scores are marked off in equal intervals like inches. This means we can directly compare children, regardless of age, because of the normalized score adjustments. The population mean is 100 with a standard deviation of 15.

Riveting Results

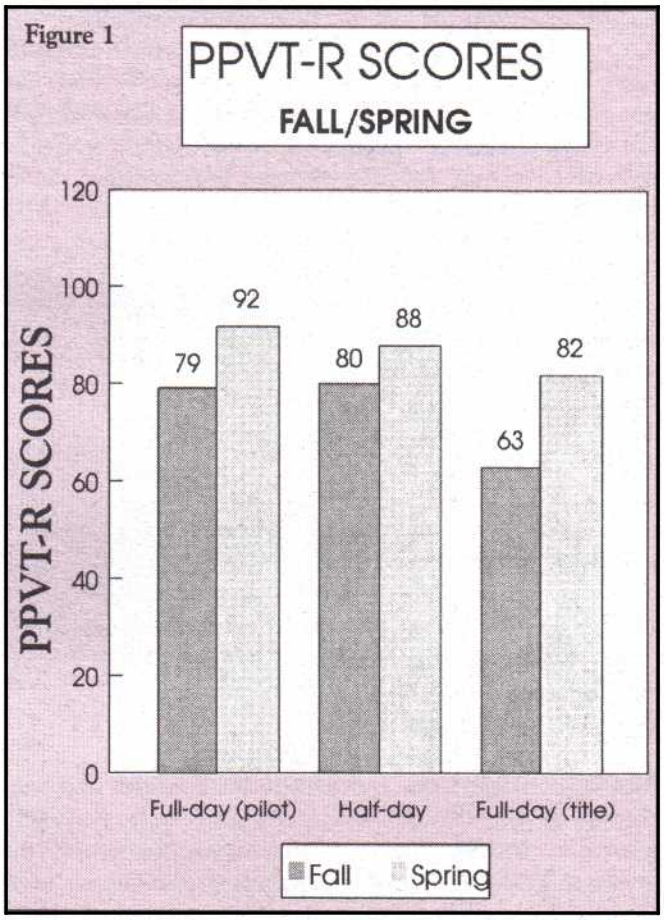
The PPVT-R was given to 440 kindergartners at the beginning of the 1997 Fall semester, and again at the end of the school year in Spring of 1998. The thirteen Title I full-day classes have 188 students and account for 43% of this sample. The ten pilot full-day general education classes comprise 36% and the five half-day general education classes 21%. Fifty-three percent of the students are female.

Figure 1 shows the results of the PPVT-R given to each kindergartner in the fall and then again in the spring.

All general education kindergartners started the school year with essentially the same entry level, 79/80. However, by the end of the year the full-day students had a significant gain of 14 points (from 79 to 92). The half-day students only gained 8 points (from 80 to 88).

The Title I full-day students were enrolled in the program for compensatory reasons. They entered school in the fall with the low entry level of 63, more than two and a half

Figure 1



standard deviations below the normalized average of 100. They exited kindergarten in the spring with a tremendous gain of over 19 points, with a score of 82, just a little over one standard deviation below the norm.

When PPVT-R measures are examined by class and by school, one finds answers to the research questions. It does not seem to make a difference whether the kindergarten class is in the elementary schools or the high schools. Individual teachers and/or schools seem to account for the variation in improvement level.

Gender did not make a difference in improvement for the general education students. However, the Title I girls improved on the PPVT-R by an average of 6 points more than the boys. (Female = 23 Male = 17) Are more linguistic opportunities available to girls in the classroom? Do teachers have more effective language activities for girls than boys? Perhaps some teachers have proven methods to share with their peers.

Another interesting finding about Title I girls is that they enter school 5 points below the boys on the PPVT-R. (Female = 60 Male = 65) Could this be because boys are given more attention and spoken to more in the home than girls?

Discussion

The substantial gain made by students enrolled in the full-day pilot or the full-day Title I kindergarten programs is

enough to support the Full-day Kindergarten Initiative.

Research shows that the earlier the brain is stimulated, the bigger the influence. The difference between the two full-day programs and the traditional half-day kindergarten is overwhelming. Indianapolis Public Schools would be well-served to fund Full-Day programs as prevention, rather than spend more money in later years for remediation.

The full-day kindergarten strategy raises the threshold for student achievement. It produces academically stronger students who are better able to compete with the "norm." Higher test scores in later years will prove the efficacy of early intervention and stimulation with well-designed, well-implemented full-day kindergarten programs.

This study provides an objective method for teacher development. Teachers who produce large gain scores are identified. These teachers can work on peer-to-peer staff development to share their classroom techniques. Teachers who need support are also identified. A mentoring program could help these teachers improve their methods.

Nancy E. Beatty, M.A., J.D.
Indianapolis Public Schools
Title I Facilitator

The school system's mandate is to educate children efficiently and effectively. Research and data analysis allows each initiative, each classroom, to become a laboratory. It allows administrators and teachers to work together to look at pedagogical programs and determine "Best Practices."

Action Plans Work!

Using the Scientific Method with precise measurement and careful evaluation, we have the tools to make more informed, defensible decisions for the welfare of our children. That is exactly what Superintendent Pat Pritchett did. He presented the results of this study to the IPS School Board. He mailed a copy of the report to every legislator in the state of Indiana who would be voting in 1999 on Indiana House Bill 1689, which sets aside funding for optional full-day kindergarten. Pritchett made these results public to prove to taxpayers and legislators that the Indianapolis Public School District is actively involved in the search for the best methods and efficient use of funds to provide quality education for its constituency.

Indiana Governor Frank O'Bannon is strongly in favor of funding optional full-day kindergarten and is an active advocate of this initiative. On January 28, 1999 the bill passed the House Education Committee with only one vote against the proposal. This proves a proactive, research-oriented approach enhances and guides pedagogy as well as public policy. The $111 million bill is Governor O'Bannon's number one priority in the General Assembly this year. We will let our readers know the outcome of the vote in the next issue.

The Indianapolis Public School District is to be commended for taking the lead to prove the efficacy of its programs. They have the courage to hold their programs up to the light of day in order to pursue their goal of providing the best possible education to their children. Their intellectual honesty will allow them to joyously proclaim their successes, and fix the programs which do not produce positive results. Proper measurement allows both accountability and the freedom to be creative and experimental. Immediate feedback produces immediate corrections.

For a copy of the complete report contact:
Nancy E. Beatty, M.A., J.D.
Indiana Public Schools Title I Facilitator
Telephone: (317) 226-3224
E-mail: nbeatty@ips.k12.in.us

Indianapolis Public Schools Superintendent Pritchett reads to kindergartners as the "Cat in the Hat."

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

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

In 1991 Donna started Meaningful Measurement, a consulting consortium for communication training, organizational development, market research, and educational assessment. Donna is Psychometrician and Director of Examination Activities for the American Society of Clinical Pathologists. Her leisure activities are swimming, yoga, and reading mystery novels.

e-mail: surgstatum@aol.com

Speak Up!

A College Competency Assessment Tool

Richard Quianthy and Deborah Hefferin

For two decades, an ongoing effort of the National Communication Association (NCA, formerly SCA) has been the identification of speaking and listening skills, the required levels of competency of those skills, and assessment of skill acquisition. This has been a priority at every educational level from kindergarten through higher education.

This project began in 1982 with the SCA study of Oral Communication Competencies Needed by Community College Graduates Entering Careers. A task force was formed to determine college student competencies in speaking and listening. Following a national survey, the task force developed a set of competencies that were reviewed and adopted by NCA's Administrative Committee in 1985. These were disseminated as the "Essential College Sophomore Speaking and Listening Competencies."

These competencies apply to all college students regardless of the institution they attend, their major, or their program of study. Whether a speech course is required, or regardless of the specific definition of that course, all students should demonstrate the same basic competencies.

A 1987 SCA Wingspread Conference further delineated the competencies and suggested teaching strategies and made suggestions for assessment of the skills. In Communication Is Life (NCA, 1990) the recognition of the dichotomy between 'what' communication is and 'how' a student communicates is stressed. The needs to measure knowledge and to evaluate performance behaviors of a competent communicator are equally important.

Evaluation and assessment of communication skills has long been central. NCA demonstrated this commitment when it created the Committee on Assessment and Testing (CAT) in 1970. This committee and subcommittees working under CAT since then have carried various projects forward. In 1990, following a national conference on assessment, NCA adopted Criteria for the Assessment of Oral Communication. The emphasis of this document is on communication as an interactive process.

There are many purposes for communication assessment. Depending on the needs of an individual institution, assessment might be used as a pretest, to place students into a course or to exempt them from it; as a posttest, to exit a course; or as large-scale assessment to identify that educational goals have been met. All of these purposes indicate a need for a test of students' communication competencies, as articulated in NCA's earlier publications.

Test Development

The current task force developed a paper-and-pencil test on knowledge, of the Sophomore Level Exit Competencies for Speaking and Listening. Specifications for this test include:

* The test is not course specific. It may be used in any course that incorporates the competencies.

* The test is not text specific. It is assumed that the competencies will be addressed in the course, although the presentation and method may differ.

* The test is not jargon specific. Care was taken to see that specialized language would not deter a student from showing understanding of a concept.

* The test is developed to show mastery of specific topics.

The task force has developed a testbank covering all of the competencies listed as a part of the sophomore competencies document. Each question is identified as fitting a specific competency. The task force members and other NCA members wrote the original set of questions This set of questions was then reviewed by members of the NCA's Communication Assessment Commission. The current testbank is the result of this entire process.

The communication practitioner has a lot of flexibility in using this document to assess competencies. They can

use the entire set of questions for an in-depth testing of the competencies, or a selection of the questions could compose a short test that would give a quick assessment of the cognitive component of the competencies. Care should be taken to be sure that all competencies are covered when choosing a selection for a short test.

The Communication Competencies

Expected Student Outcomes for Speaking and Listening: Basic Communication Course and General Education

The following student outcomes represent some of the expectations for students taking a basic communication course and/or participating in the general education requirements of a school. Basic course or general education students need speaking and listening skills that will help them succeed in future course and on the job. They need to be able to construct and deliver messages and listen with literal and critical comprehension. The basic course can provide knowledge of effective communication techniques, an arena for developing and practicing skills, and positive feelings about communicating in the future. Instructors and administrators could use some or all of the expected student outcomes to inform the design of a basic communication course. Academic institutions could use some of all of the outcomes to describe campus expectations for students in regard to the general education curriculum (Rosenbaum, 1994).

I. SpeakingCmpetencies

Speaking is the process of transmitting ideas and information orally in a variety of situations. Effective oral communication involves generating messages and delivering them with attention to vocal variety, articulation, and nonverbal signals.

In order to be a COMPETENT SPEAKER, a person must be able to compose a message and provide ideas and information suitable to the topic, purpose, and audience. The COMPETENT SPEAKER must also be able to transmit the message by using delivery skills suitable to the topic purpose and audience. In addition, the COMPETENT SPEAKER must be able to transmit messages using interpersonal skills suitable to the context and the audience.

II. Listening Ccompetencies

Listening is the process of receiving, constructing meaning from, and responding to spoken and/or nonverbal messages. People listen in order to comprehend information, critique and evaluate a message, show empathy for the feelings expressed by others, or appreciate a performance. Effective listening includes both literal and critical comprehension of ideas and information transmitted in oral language.

Data Analysis

Validity

The Cognitive Test for College Level Communication Competencies meets the criteria for face and content validity. The multiple-choice questions were carefully designed and refined by communication experts. The test was sent for review to colleagues who gave feedback and suggestions. The final version was then sent to members of the NCA Committee on Assessment and Testing. It was on the agenda at the 1997 Chicago Convention for the CAT business meeting where it was discussed thoroughly and given approval for further testing.

Data

This test was given as a final exam for speech communication classes in the 1997 Fall semester at Broward Community College in Florida. Some classes were straight public speaking; others were a hybrid of communication theory and public speaking. One hundred forty-six students took the test.

The students are predominantly white (93). Twenty-one identify themselves as African American, fourteen students are Hispanic. Two-thirds of this group of students is female.

Three-quarters of this group are traditional students between the ages of 17 and 22. Twenty-one are between 23 and 29 years old. Eleven are over thirty.

Results

Item Analysis

The first thing that is done in a Rasch analysis is to "test the test"— to examine the items on the survey to make sure they are creating a valid ruler to measure the variable.

Do the items cover the range of the variable? It is not useful if everything is bunched up together. It would be like giving a test with only simple addition problems. We would not know whether the person could perform other mathematical functions — only whether he or she could add. So, too, with "speech communication competency." If we have a range of easier to harder items, then we have an indication of the level of a person's speech communication ability.

Item Fit

Do all of the items "fit"? Are we measuring what we think we are measuring? Which items, if any, need to be rewritten for future surveys? Checking allows us to be sure we are only measuring one thing at a time.

The Communication Competency Assessment passed all tests with flying colors. The items fit and have a wide range of difficulty. This means we have developed a calibrated in-

strument that measures what it is designed for and can be used to examine student competency.

Logits

The units of measure are called "logits" and each logit has 100 points. When reading this report, all numbers are directly comparable to each other. The results for each item or person are in the same units of measure. Thus we can compare students from this year to the next, or one class to another common frame of reference. This gives a benchmark, which can be used to compare with future performance, and help in establishing goals for improvement.

Measurement Scale

The scale has been calibrated so the origin, or balance point, is 10.00. That means a person who is "average" in ability/competence, or an item which is of "average" difficulty, has a measure of 10.00. An item calibrated at 10.00 has a 50/50 chance of being answered correctly by a person who has a 10.00 measure of ability. The lower the number, the less ability the person has, or the easier the item is to answer correctly. Measures higher than 10 indicate more ability/ competence than that of the "average" person, or an item that is harder than average.

Person Summary Statistics

The average raw score is 46.7 out of 75 items. The average person measure is 10.67 The model error is .27. The standard deviation is .60. The separation is 1.88 and reliability is .78.

These statistics mean this group of students is fairly competent as a whole: .67 logits above the mythical average. Each student measure is accurate to within about a quarter of a logit, or .27 logit.

The standard deviation shows the shape, or spread, of the distribution. Ninety-five percent of the students will be within two standard deviations up or down from the average measure of 10.67. In other words, 95% of these students will have a measure within the range of 11.87 and 9.47.

The person separation of 1.88 is not terribly high, and person reliability is .78. This means some people are similar in their competence. The reliability tells us that 78% of the time the person measures will give the same order for person competence. In other words, 22% of the variance in person measures is due to estimation error.

Item Summary Statistics

The average item measure is 10.00. The model error is .21 The standard deviation is 1.11. The separation is 5.11 and the item reliability is .96.

The center point for item difficulty is set at 10.00. Items higher than 10.00 are more difficult than average and items lower than 10.00 are easier. Item calibrations are accurate within an error of 21 points, or about a fifth of a logit.

Item separation is very high at 5.11, and item reliability is.96. That means only 4% of the variance in the item calibration is due to estimation error. This excellent reliability allows us to have confidence in the items defining "Speech Communication Competence."

The Person - Item Map (**Figure 1**)gives a visual report of the results. Along the left side is the logit ruler which measures the placement of the persons and items. Remember, the important, unique feature of this method of analysis is that each facet is measured independently. The higher the measure (the placement on the page), the more the person's ability or the more difficult the item.

The map shows the students' measures are distributed in a normal curve. Each # stands for two cases; a . is an individual. The person measure of ability is from a low of 8.29 to a high of 12.14, a range of almost 400 points.

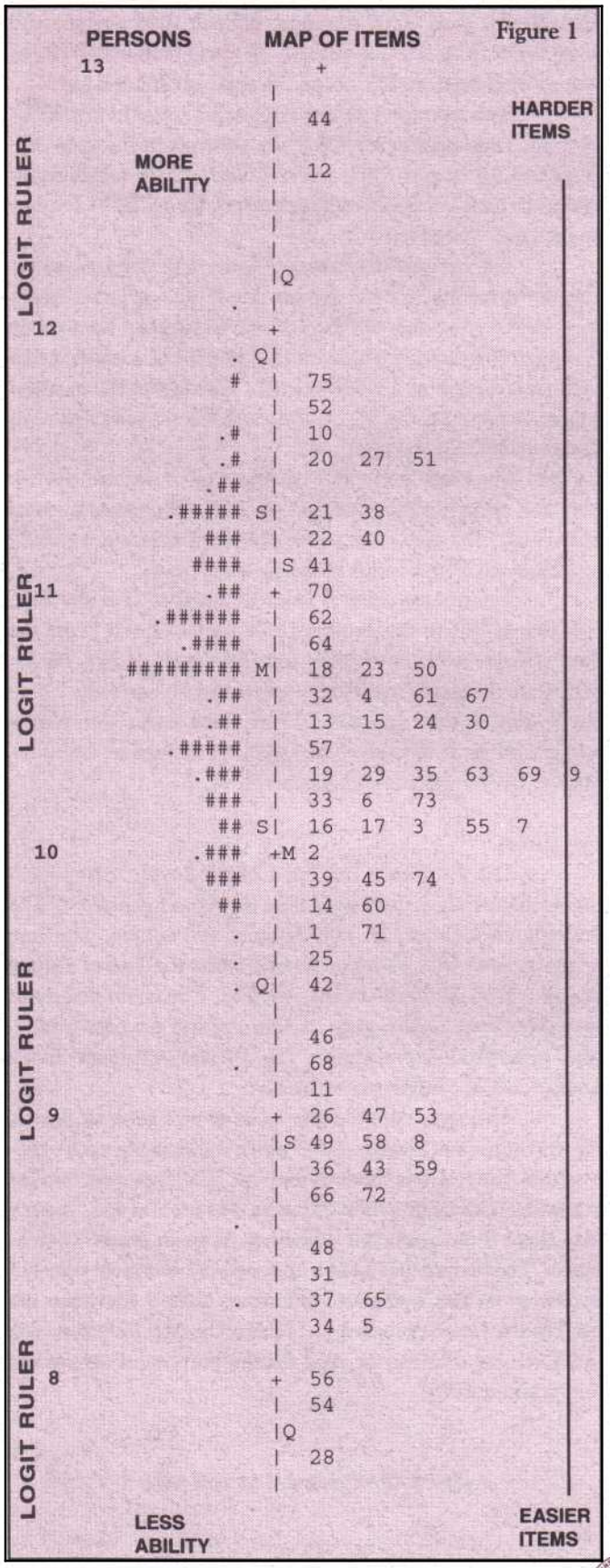
Items range about 5 logits in levels of difficulty from very easy at 7.65 to very hard at 12.76. Items 44 and 12 are the most difficult items at the very top of the page. Items 28, 54, and 56 are the easiest items for people to answer correctly. There are no gaps in item difficulty. These items cover a very wide range, and provide a good yardstick to determine cognitive Speech Communication Competency.

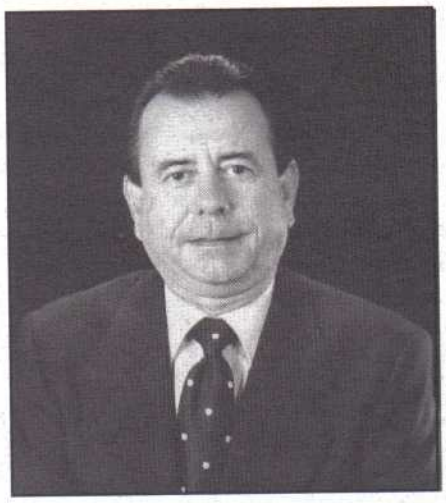
Discussion

The Cognitive Test For College Level Communication Competencies Assessment is an excellent instrument. The multiple-choice items are well-designed and refined. The item analysis shows that all items fit well along the line of inquiry and provide a definition of the variable. The calibrated items spread out over a wide range of difficulty and can clearly identify a person's ability measure. The .96 item reliability allows a great deal of confidence in this test.

Although some demographic groups such as Asians, American Indian, Hispanic, and older students are under-represented, the test was checked for bias. The data were divided according to demographics: ethnicity, gender, and age. In each case there is no significant difference in mean person competence. The items behaved in a uniform way and there does not appear to be any systematic difference in how the items are used by the various subgroups. Thus far the data indicate a fair, unbiased test that can be used for the purpose of assessment and accountability.

(Speak Up! Continued on next page)





Richard Quianthy

Richard Quianthy served as Project Director for the development of this test. He is a Professor of Communication at Broward Community College, Ft. Lauderdale, where he teaches Public Speaking and is the recipient of an Endowed Teaching Chair. Richard has been involved with the development and assessment of speaking and listening competencies since 1983. He is currently serving as the Vice-chair of the National Communication Association's Communication Assessment Commission. He has been recognized by NCA as an Outstanding Community College Educator and for "Outstanding Contributions to the Field of Assessment."

Deborah Hefferin

Deborah Hefferin also is a Professor of Communication at Broward Community College and holds an Endowed Teaching Chair. Since 1982 she has worked on various projects developing communication standards and competencies at the K-12 and college level. Her interests include taking the standards and working to develop teaching activities and assessment instruments to complete the project. Deborah recently served on the NCA Educational Policies Board and received the 1996 NCA Community College Educator Award.

Measuring Change In Efficacy

Everett V. Smith, Jr.
University of Illinois at Chicago
Kimberly A. Lawless
Utah State University
Leslie Curda and Steven Curda
University of Toledo

"I think I can, I think I can"

(Piper, W. (1954). The Little Engine That Could).

As we all remember, the little engine succeeded in climbing the hill. Not because it knew it could succeed based on past performance, but because it believed it could succeed. This self-belief in one's ability to perform a specific task is known as self-efficacy (Bandura, 1986). It is a person's "I can" or "I cannot do" belief. It is not concerned with the skills one has, but with the judgments of what one can do with the skills one possesses.

Why Measure Self-efficacy?

A primary purpose of training programs is to convey a body of knowledge which can then be applied in the future. How effective is instruction in conveying this body of knowledge? The typical evaluation of instruction only assesses acquisition of knowledge while overlooking self-efficacy. Self-efficacy influences persistence and motivation, important outcomes to any training program. In conjunction with achievement data, self-efficacy measures can serve as an important part of program planning and evaluation (Owen, 1991), (indicating areas where individuals do not possess sufficient confidence in their ability to perform specific tasks, either prior to, during, or following a program of instruction.) Perceived weaknesses can suggest more efficient instruction. As Lusardi and Smith (1997) discuss, self-efficacy measures can be better indicators of use of recently acquired knowledge than outcome measures. Evidence that knowledge has been learned (outcome measures) is not evidence that knowledge will be applied. A training program is limited if alteration of behavior is achieved but the learner is not endowed with confidence to engage in the behavior at some future point in time.

Background

We evaluated change in self-efficacy for a group of undergraduate preservice teachers (n=48) enrolled in an instructional technology course. One of the issues facing the educational field today is the incorporation of new teacher competencies into existing requirements for teacher certification (e.g., ISTE, NCATE). Many of these competencies deal with teachers' ability to utilize a variety of technologies. As these new competencies become more common and teacher preparation programs become accountable for graduating students with these skills, courses must be developed and evaluations conducted to ensure that competencies are being met. As these courses are developed, teacher educators must deal with issues that may interfere with their students' willingness to engage in new experiences with technology. Preservice teachers' efficacy in the use of technologies is a key factor that will influence whether they are willing to integrate technology into the curriculum.

The Classroom Technology Questionnaire (CTQ) was designed to assess students' self-efficacy in implementing instructional technologies. The CTQ consists of 14 items, each focusing on a different form of instructional media. For each type of media, a definition was provided to help all respondents to respond from a common frame of reference. For each form of media and item stem, respondents were asked the following question: "Imagine you are teaching RIGHT NOW. How skillful do you feel about using this type of media in your classroom RIGHT NOW?" Responses were collected on a 7-point Likert type scale, with poles labeled NOT AT ALL SKILLFUL (1) and EXTREMELY SKILLFUL (7).

The evaluation

Program evaluations typically use aggregate data and assess the impact of a program using a sample dependent t-test. This results in an evaluation of whether the group mean is significantly different from pretest to posttest. For diagnostic pur-

poses, there is a need to be able to locate individuals who are different from the group both prior to instruction and upon completion of the course. Further, for valid pre-post interpretations, the potential problem associated with the functioning of the items, which may be interpreted differently at each measurement occasion by the respondents, needs to be investigated.

Results

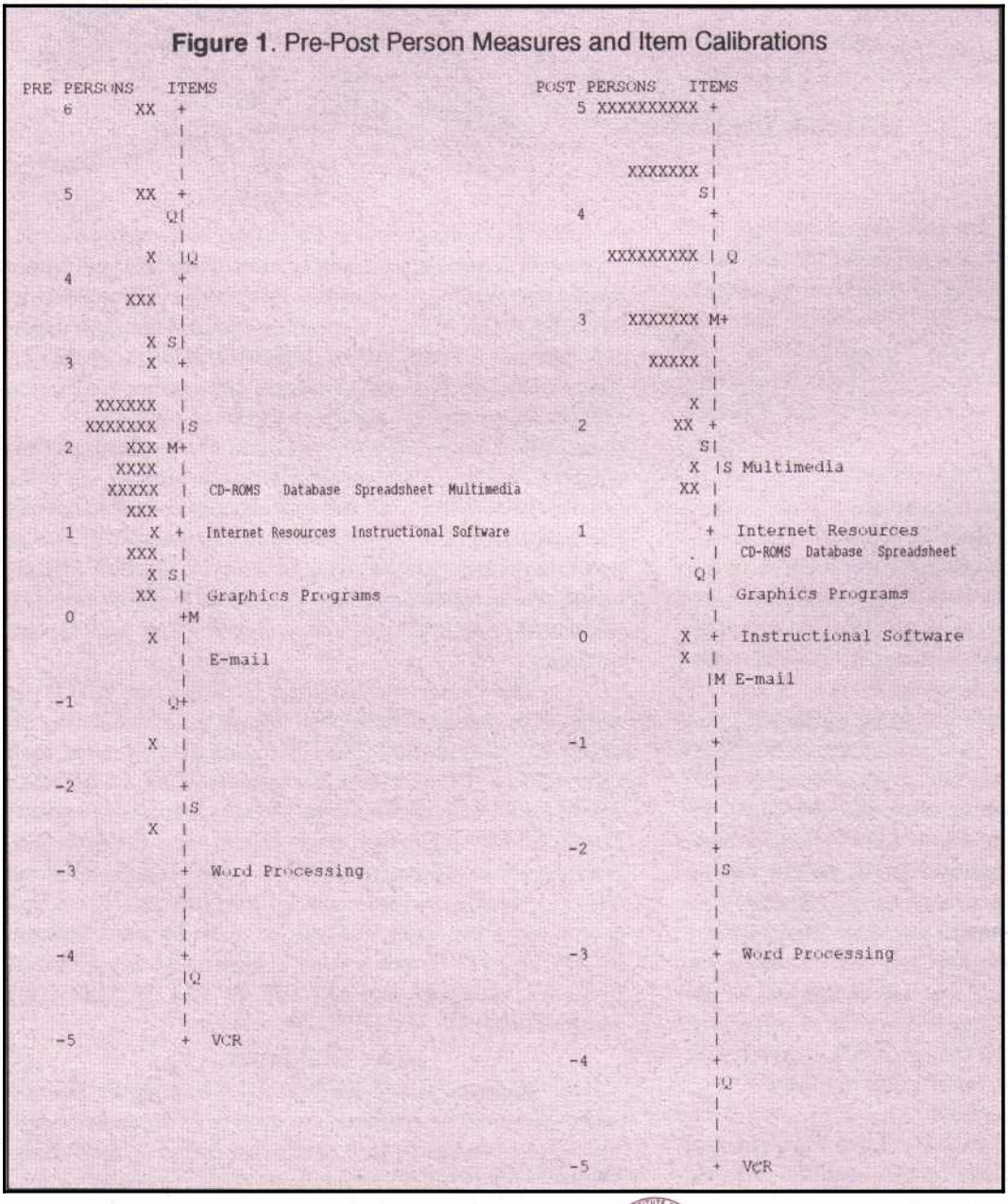
The results presented are based on ten, of the fourteen, items that were found to fit the Rasch Rating Scale Model. It was also found that a 4-point scale better represents the data than the original 7-point scale.

The traditional method of assessing change using a dependent t-test demonstrated statistically significant gains for the group from pretest to posttest (**Table 1**). This method of program evaluation has two limitations. First, changes in the underlying variable are not investigated. If the variable being measured is not the same from pretest to posttest, evaluation of change is meaningless (see Wright, 1996). Second, rather than concentrating on group differences, it would be of greater value to see which individuals demonstrated statistically significant gains or losses.

Table 1
Results of pre-post analysis at the group level

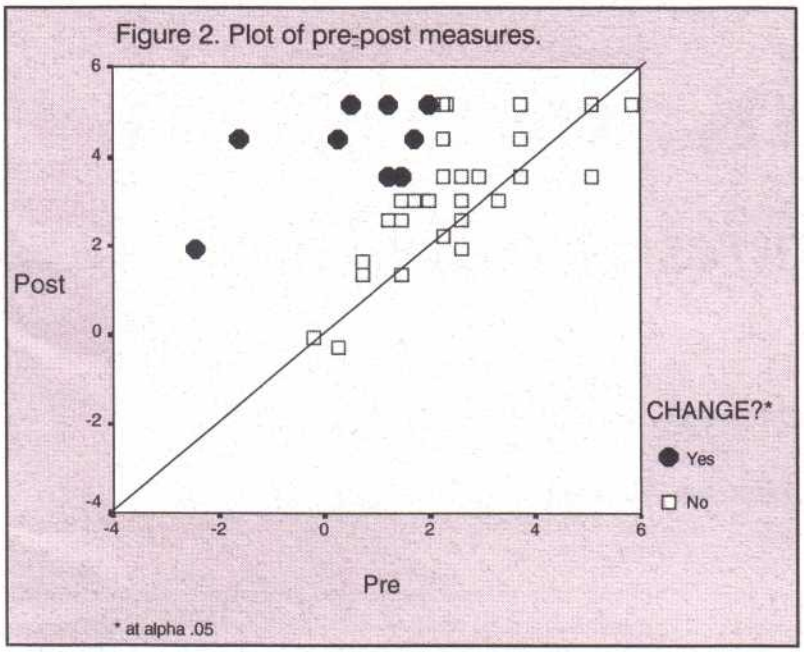
	Mean	SD	Mean Difference	SD	t	p
Pre measure	2.08	1.67	1.36	1.66	5.29	.001
Post measure	3.44	1.48				

Figure 1. Pre-Post Person Measures and Item Calibrations



Fortunately, Rasch measurement can be used to address both of these deficiencies. **Figure 1** compares the variable being measured at pretest and posttest. Notice that several of the items maintain their location on the variable, indicating stability (invariance) of the item calibrations. This type of evidence is required in order to make valid pre-post comparisons. **Figure 1** also displays the shift in person measures (the shift in the group mean labeled 'M'), and pictorially represents the results of the dependent t-test (Note **Figure 1** represents calibration of all available data from pretest (n=48) and posttest (n=46) administrations, while the dependent t-test is based on complete data only (n=42)).

Rasch measurement also produces standard errors for



each measure. This distinctive advantage over Classical Test Theory allows for the statistical comparisons of pre-post scores at the individual rather than group level. Using this information, one is able to target individuals that did not display statistically significant gains in self-efficacy and those that demonstrated reductions for further self-efficacy enhancing activities. **Figure 2** demonstrates analysis of change at the individual level. Darkened squares above the identity line indicate statistically significant gains for those individuals from pretest to posttest. Nine students demonstrated statistically significant gains, none statistically significant reductions (fortunately). This information could prove of great value in evaluating a current course or planning a future course, if follow-up procedures are undertaken to investigate how and why the program benefited several individuals while seemingly not affecting others.

References

Bandura, A. (1986). Social Foundations of Thought and Actions: A Social Cognitive Theory. New Jersey: Prentice-Hall, Inc.

Lusardi, M.M., & Smith, Jr., E.V. (1997). Development of a scale to assess concern about falling and applications to treatment programs. Journal of Outcome Measurement, 1, 34-55.

Owen, S.V. (1991, April). Using self-efficacy in program evaluation. Paper presented at The Training Excellence: A Conference on training practitioners in the drugs/alcohol field, Canterbury, England.

Piper, W. (1954). The Little Engine that Could. New York: Platt and Munk.

Wright, B.D. (1996). Time 1 to Time 2 Comparisons. Rasch Measurement Transactions, 10(1), 478-479.

Kimberly A. Lawless, Assistant Professor of Instructional Technology, Utah State University. Teaching areas include technology in education, learning theory, research methods, and product evaluation. Research interests include reading comprehension, hypertext processing, teacher beliefs, and instructional design practice. Hobbies include hiking, refinishing furniture, and spending time with family and friends.

L. K. Curda is an Assistant Professor and Program Coordinator of Educational Technology at the University of Toledo. Her research activities include collaborating with university and public school colleagues to investigate groups of ecological variables to explicate the influences they have on teachers' expectancies and goals and their performance in the classroom related to technology integration in content areas. Much of her early research has been focused on developing valid and reliable measures for investigating variables associated with change management and diffusion of innovations that can be used to evaluate the effectiveness of preparation programs in developing motivated teachers, and to profile a school's practicing teachers in order to develop and implement needed interventions to increase teacher motivation for and implementation of technology-related initiatives. Email: lcurda@utoledo.edu

"When the Rasch model is intended to hold because of its special measurement properties, failure of the data to conform to the model implies further work on the substantive problem of scale construction, not on the identification of a more complex model that might account for the data."

David Andrich
in *Rasch Models for Measurement*.
1988. p.86. Newbury Park, CA: Sage.

One Fish, Two Fish

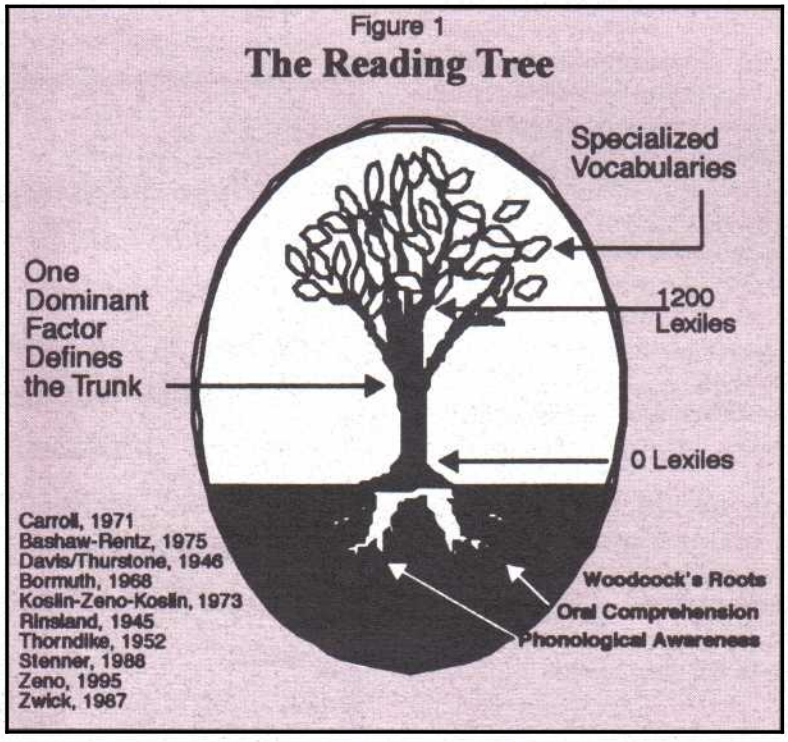
Rasch Measures Reading Best

Benjamin D. Wright
and
A. Jackson Stenner

Think of reading as the tree in **Figure 1**. It has roots like oral comprehension and phonological awareness. As reading ability grows, a trunk extends through grade school, high school, and college branching at the top into specialized vocabularies. That single trunk is longer than many realize. It grows quite straight and singular from first grade through college.

Reading has always been the most researched topic in education. There have been many studies of reading ability, large and small, local and national. When the results of these studies are reviewed, one clear picture emerges. Despite the 97 ways to test reading ability, many decades of empirical data document definitively that no researcher has been able to measure more than one kind of reading ability (Mitchell, 1985). This has proven true in spite of intense interest in discovering diversity. Consider three examples: the 1940s Davis Study, the 1970s Anchor Study, and six 1980s and 1990s ETS studies.

Figure 1
The Reading Tree



Davis - 1940s

Fred Davis went to a great deal of trouble to define and operationalize nine kinds of reading ability (1944). He made up nine different reading tests to prove the separate identities of his nine kinds. He gave his nine tests to hundreds of students, analyzed their responses to prove his thesis, and reported that he had established nine kinds of reading. But when Louis Thurstone reanalyzed Davis' data (1946), Thurstone showed conclusively that Davis had no evidence of more than one dimension of reading.

Anchor Study - 1970s

In the 1970s, worry about national literacy moved the U.S. government to finance a national Anchor Study (Jaeger, 1973). Fourteen different reading tests were administered to a great many children in order to uncover the relationships among the 14 different test scores. Millions of dollars were spent. Thousands of responses were analyzed. The final report required 15,000 pages in 30 volumes — just the kind of document one reads overnight, takes to school the next day and applies to teaching (Loret et al., 1974). In reaction to this futility, and against a great deal of proprietary resistance, Bashaw and Rentz were able to obtain a small grant to reanalyze the Anchor Study data (1975, 1977). By applying new methods for constructing

objective measurement (Wright and Stone, 1979), Bashaw and Rentz were able to show that all 14 tests used in the Anchor Study — with all their different kinds of items, item authors, and publishers — could all be calibrated onto one linear "National Reference Scale" of reading ability.

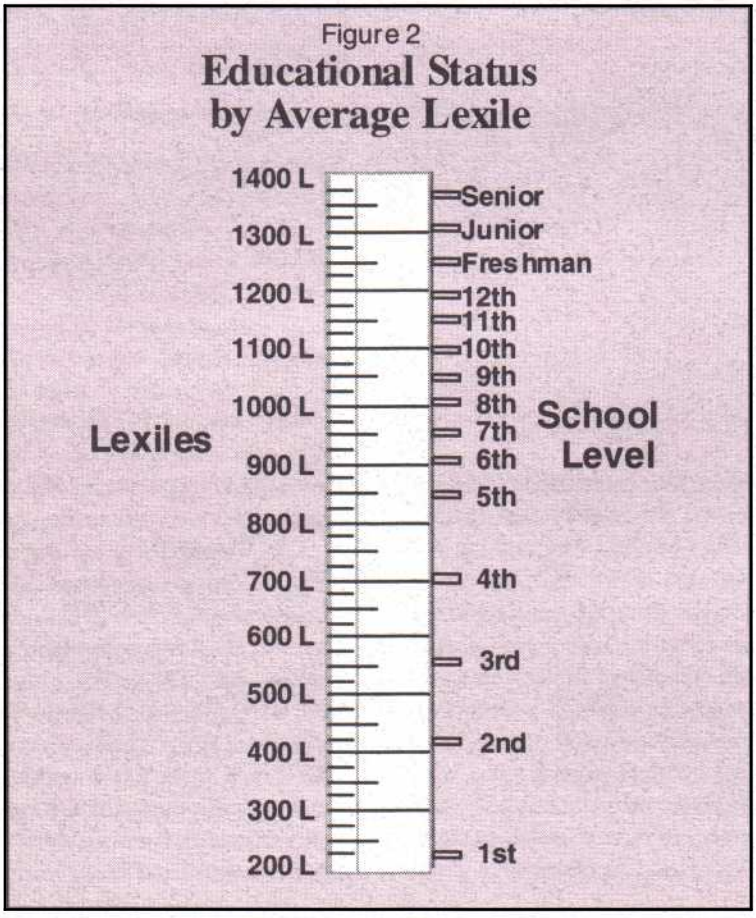
The essence of the Bashaw and Rentz results can be summarized on one easy-to-read page (1977) — a bit more useful than 15,000 pages. Their one-page summary shows how every raw score from the 14 Anchor Study reading tests can be equated to one linear National Reference Scale. Their page also shows that the scores of all 14 tests can be understood as measuring the same kind of reading on one common scale. The Bashaw and Rentz National Reference Scale is additional evidence that, so far, no more than one kind of reading ability has ever been measured. Unfortunately, their work had little effect on the course of U.S. education. The experts went right on claiming there must be more than one kind of reading — and sending teachers confusing messages as to what they were supposed to teach and how to do it.

ETS Studies - 1980s and 1990s

In the 1980s and 1990s, the Educational Testing Service (ETS) did a series of studies for the U.S. government. ETS (1990) insisted on three kinds of reading: prose reading, document reading, and quantitative reading. They built a separate test to measure each of these three kinds of reading —greatly increasing costs. Versions of these tests were administered to samples of school children, prisoners, young adults, mature adults, and senior citizens. ETS reported three reading measures for each person and claimed to have measured three kinds of reading (Kirsch & Jungeblut, 1986). But reviewers noted that, no matter which kind of reading was chosen, there were no differences in the results (Kirsch & Jungeblut, 1993, 1994; Reder, 1996; Zwick, 1987). When the relationships among reading and age and ethnicity were analyzed, whether for prose, document, or quantitative reading, all conclusions came out the same.

Figure 2

Educational Status by Average Lexile



Later, when the various sets of ETS data were reanalyzed by independent researchers, no evidence for three kinds of reading measures could be found (Bernstein, & Teng, 1989; Reder, Rock and Yamamoto, 1994; 1996; Salganik and Tal, 1989; Zwick, 1987). The correlations among ETS prose, document, and quantitative reading measures ranged from 0.89 to 0.96. Thus, once again and in spite of strong proprietary and theoretical interests in proving otherwise, nobody had succeeded in measuring more than one kind of reading ability.

Lexiles

Figure 2 is a reading ruler. Its Lexile units work just like the inches. The Lexile ruler is built out of readability theory, school practice, and educational science. The Lexile scale is an interval scale. It comes from a theoretical specification of a readability unit that corresponds to the empirical calibrations of reading test items. It is a readability ruler. And it is a reading ability ruler.

Readability formulas are built out of abstract characteristics of language. No attempt is made to identify what a word or sentence means. The idea is not new. The Athenian Bar Association used readability calculations to teach lawyers to write briefs in 400 B.C. (Chall, 1988; Zakaluk and Samuels, 1988). According to the Athenians, the ability to read a passage was not the ability to interpret what the passage was about. The ability to read was just the ability to read. Talmudic teachers who wanted to regularize their students' studies, used readability measures to divide the Torah readings into equal portions of reading difficulty in 700 A.D. (Lorge). Like the Athenians, their concern in doing this was not with what a particular Torah passage was about, but rather the extent to which passage readability burdened readers.

In the twentieth century, every imaginable structural characteristic of a passage has been tested as a potential source for a readability measure: the number of letters and syllables in a word; the number of sentences in a passage; sentence length; balances between pronouns and nouns, verbs and

prepositions (Stenner, 1997). The Lexile readability measure uses word familiarity and sentence length.

Lexile Accuracies

Table 1 lists the correlations between readability measures from the ten most studied readability equations and student responses to different types of reading test items. The columns of Table 1 report on five item types:

Lexile Slices;
SRA Passages;
Battery Test Sentences;
Mastery Test Cloze Gaps;
Peabody Test Pictures.

The item types span the range of reading comprehension items. The numbers in the table show the correlations between theoretical readability measures of item text and empirical item calibrations calculated from students' test responses. Consider the top row. The Lexile readability equation predicted

Table 1
Correlations between
Empirical & Theoretical Item Difficulties

Ten Readability Equations	Five Test Item Types				
	Lexile Slice	SRA Passage	Battery Sentence	Mastery Cloze	Peabody Picture
Lexile	.90	.92	.85	.74	.94
Flesch	.85	.94	.85	.70	.85
ARI	.85	.93	.85	.71	.85
FOG	.85	.92	.73	.75	.85
Powers	.82	.93	.83	.65	.74
Holquist	.81	.91	.81	.84	.86
Flesch-1	.79	.92	.81	.61	.69
Flesch-2	.75	.87	.70	.52	.71
Coleman	.74	.87	.75	.75	.83
Dale-Chall	.76	.92	.82	.73	.67

Adapted from Stenner, 1997

how difficult Lexile slices would be for persons taking a Lexile reading test at a correlation of 0.90, the SRA passage at 0.92, the Battery Sentence at 0.85, the Mastery Cloze at 0.74, and the Peabody Picture at 0.94 (Stenner, 1996). With the exception of the cloze items, these predictions are nearly perfect. Also note that the simple Lexile equation, based only on word familiarity and sentence length, predicts empirical item responses as well as any other readability equation— no matter how complex. **Table 1** documents, yet again that one, and only one, kind of reading is measured by these reading tests. Were that not so, the array of nearly perfect correlations could not occur. **Table 1** also shows that we can have a useful measurement of text readability and reader reading ability on a single reading ruler!

An important tool in reading education is the basal reader. The teaching sequence of basal readers records generations of practical experience with text readability and its bearing on student reading ability. **Table 2** lists the correlations between Lexile Readability and Basal Reader Order for the eleven basal readers most used in the United States. Each series is built to mark out successive units of increasing reading difficulty. Ginn has 53 units — from book 1 at the easiest to book 53 at the hardest. HBJ Eagle has 70 units. Teachers work their students through these series from start to finish. **Table 2**

Table 2
Correlations between
Basal Reader Order & Lexile Readability

Basal Reader Series	Basal Units	r	R	R'
Ginn	53	.93	.98	1.00
HBJ Eagle	70	.93	.98	1.00
SF Focus	92	.84	.99	1.00
Riverside	67	.87	.97	1.00
HM (1983)	33	.88	.96	.99
Economy	67	.86	.96	.99
SF Amer Trad	88	.85	.97	.99
HBJ Odyssey	38	.79	.97	.99
Holt	54	.87	.96	.98
HM (1986)	46	.81	.95	.97
Open Court	52	.54	.94	.97

Adapted from Stenner, 1997

r = raw R = corrected for attenuation R' = corrected for attenuation and range restriction

shows that the correlations between Lexile measures of the texts of these basal readers and their sequential positions from easy to hard are extraordinarily high. In fact, when corrected for attenuation and range restriction, these correlations approach perfection (Stenner, 1997)

Each designer of a basal reader series used their own ideas, consultants, and theory to decide what was easy and what was hard. Nevertheless, when the texts of these basal units are Lexiled, these Lexiles predict exactly where each book stands on its own reading ladder — more evidence that, despite differences among publishers and authors, all units end up benchmarking the same single dimension of reading ability.

Finally there are the ubiquitous reading ability tests administered annually to assess every student's reading ability. **Table 3** shows how well theoretical item text Lexiles predict actual readers' test performances on eight of the most popular reading tests. The second column shows how many passages from each test were Lexiled. The third column lists the item type. Once again there is a very high correlation between the difficulty of these items as calculated by the entirely abstract Lexile specification equation and the live data produced by students answering these items on reading tests. When we correct for attenuation and range restriction, the correlations are just about perfect. Only the Mastery Cloze test, well-known to be idiosyncratic, fails to conform fully.

What does this mean? Not only is only one reading ability being measured by all of these reading comprehension

READING RULER

Table 3
Correlations between
Passage Lexiles & Item Readabilities

Tests	Passages Analyzed	Item Type	r	R	R'
SRA	46	Passage	.95	.97	1.00
CAT-E	74	Passage	.91	.95	.98
CAT-C	43	Passage	.83	.93	.96
CTBS	50	Passage	.74	.92	.95
NAEP	70	Passage	.65	.92	.94
Lexile	262	Slice	.93	.95	.97
PIAT	66	Picture	.93	.94	.97
Mastery	85	Cloze	.74	.75	.77

Adapted from Stenner, 1997

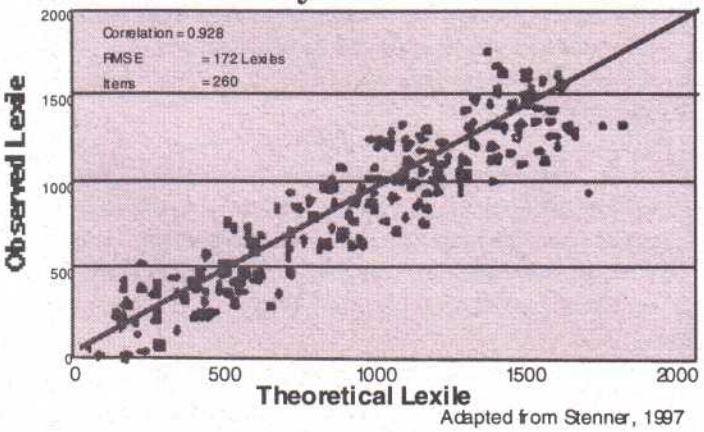
r = raw R = corrected for attenuation R' = corrected for attenuation and range restriction

tests, but we can replace all the expensive data used to calibrate these tests empirically with one formula — the abstract Lexile specification equation. We can calculate the reading difficulty of test items by Lexiling their text without administering them to a single student!

Figure 3 puts the relationship between theoretical Lexiles and observed item difficulties into perspective. The uncorrected correlation of 0.93, when disattentuated for error and corrected for range restrictions, approaches 1.00. The Lexile equation produces an almost perfect correlation between theory and practice.

Figure 3 shows the extent to which idiosyncratic variations in student responses and item response options enter the process. Where does this variation come from? Item response options have to compete with each other or they do not work. But there has to be one correct answer. Irregularity in the composition of multiple-choice options, even when they are reduced to choosing one word to fill a blank, is unavoidable. What the item writer chooses to ask about a passage and the options

Figure 3
Theory into Practice



Adapted from Stenner, 1997

they offer the test taker to choose among are not only about reading ability. They are also about personal differences among test writers.

There are also variations among test takers in alertness and motivation that disturb their performances. In view of these unavoidable contingencies, it is surprising that the correlation between Lexile theory and actual practice is so high. How does this affect the measurement of reading ability? The root mean square measurement error for a one-item test would be about 172 Lexiles. What are the implications of that much error? The distance from First Grade school books to Second Grade school books is 200 Lexiles. So we would undoubtedly be uneasy with measurement errors as large as 172 Lexiles. However, when we combine the responses to a test of 25 Lexile items, the measurement error drops to 35 Lexiles. And when we use a test of 50 Lexile items, the measurement error drops to 25 Lexiles — one-eighth of the 200 Lexile difference between First and Second Grade books. Thus, when we combine a few Lexile items into a test, we get a measure of where a reader is on the Lexile reading ability ruler, precise enough for all practical purposes. We do not plumb their depths of understanding. But we do measure their reading ability.

Sources

Bashaw, W.L. & Rentz, R.R. (1975). Equating Reading Tests with the Rasch Model, v1: Final Report & vol1 & vol2: Technical Reference Tables. Final Report of U.S. Department of Health, Education, and Welfare Contract OEC-O72-5237. Athens, GA: The University of Georgia. (ERIC Document Reproduction Nos. ED 127 330 & ED 127 331.)

Bashaw, W.L. & Rentz, R.R. (1977). The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, 14:161-179.

Bernstein, I.H., & Teng, G. (1989). Factoring Items and Factoring Scales are Different: Spurious Evidence for multidimensionality due to Item Categorization. Psychological Bulletin, 105(3):467-477.

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

Carroll, J.B., Davies, P. & Richmond, B. (1971). The Word Frequency Book, Boston: Houghton Mifflin,.

Campbell, A., Kirsch, I.S., & Kolstad, (1992). A. Assessing Literacy: The Framework for the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Chall, J.S. (1988). "The Beginning Years." In B.L. Zakaluk and S.J. Samuels (Eds.), Readability: Its Past, Present and Future, Newark, DE: International Reading Association.

Davis, Fred. (1944). Fundamental Factors of Comprehension in Reading, Psychometrika, 9:185-197.

Educational Testing Service. (1990). ETS Tests of Applied Literacy Skills. NY: Simon & Schuster Workplace Resources.

Jaeger, R.M. (1973). The National Test Equating Study in Reading (The Anchor Test Study). Measurement in Education, 4:1-8.

Kirsch, I.S., Jungeblut, A. & Campbell, A. The ETS Tests of Applied Literacy. Princeton, MJ: Educational Testing Service, 1991.

Kirsch, I.S., Jungeblut, A., Jenkins, L., & Kolstad, A. Adult Literacy in America: A First look at the Results of the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics, U.S. Department of Education, 1993.

Kirsch, I.S., Jungeblut, A., & Mosenthal, P.B. (1994). "Moving Toward the Measurement of Adult Literacy," paper presented at the March NCES Meeting, Washington, DC, 1994.

Loret, P.G.; Seder, A.; Bianchini, JC. & Vale, C.A. (1974). Anchor Test Study Final Report: Project Report and vols. 1-30. Berkeley, CA: Educational Testing Service. (ERIC Document Nos. Ed 092 601 - ED 092 631.

Lorge, I. (1939). Predicting Reading Difficulty of Selections for

School Children. Elementary English Review, 16:229-233.

Mitchel, J.V. (1985). The Ninth Mental Measurements Yearbook. Lincoln, NE: University of Nebraska Press.

Reder, Stephen, (1996). "Dimensionality and Construct Validity of the NALS Assessment," in M.C. Smith (Ed.) Literacy for the 21st Century: Research, Policy and Practice, Greenwood Publishing, in Press.

Rentz, R.R. & Bashaw, (1975) W.L. Equating Reading Tests with the Rasch Model, v1: Final Report & vol1 & vol2: Technical Reference Tables. Final Report of U.S. Department of Health, Education, and Welfare Contract OEC-O72-5237. Athens, GA: The University of Georgia. (ERIC Document Reproduction Nos. ED 127 330 & ED 127 331.

Rentz, R.R. & Bashaw, W.L. (1977). The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, 14:161-179.

Rinsland, H.D. A Basic Vocabulary of Elementary School Children, 1945.

Rock, D.A., & Yamamoto, K. (1994). Construct Validity of the Adult Literacy Subscales. Princeton, NJ: Educational Testing Service.

Salganik, L.H., & Tal, J. (1989). A Review and Reanalysis of the ETS/NAEP Young Adult Literacy Survey. Washington, DC: Pelavin Associates.

Stenner, A.J., & Burdick, D.S. (1997) "The Objective Measurement of Reading Comprehension: In Response to Technical Questions Raised by the California Department of Education Technical Study Group." Durham, NC: Metametrics.

Thorndike, E. L. & Lorge, I. (1952). The Teacher's Word Book of 30,000 Word.

Thurstone, L.L. (1946). "Note on a Reanalysis of Davis' Reading Tests," Psychometrika, v11, n2, 185ff.

Woodcock, R.W., Woodcock (1974). Reading Mastery Tests. Circle Pines, MN: American Guidance Service.

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

Zakaluk and S.J. Samuels (Eds.) (1988). Readability: Its Past, Present and Future, Newark, DE: International Reading Association.

Zeno, S.M., Ivens, S.H., Millard, R.T. & Davvuri, Raj. (1995). The Educators Word Frequency Guide, Touchstone.

Zwick, R. (1987). Assessing the Dimensionality of the NAEP Reading Data, Journal of Educational Measurement, 24:293-308.

**The authors are grateful to Ed Bouchard for helping with this report.*

A. Jackson Stenner, Ph.D.

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

Jack Stenner is also former Chairman and co-founder of National Technology Group, a 700-person firm specializing in computer networking and systems integration which was sold to VanStar Corporation in December 1996. He holds a Ph.D degree from Duke University and Bachelor degrees in Psychology and Education from the University of Missouri.

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

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

Lexile Perspectives

Benjamin D. Wright and A. Jackson Stenner

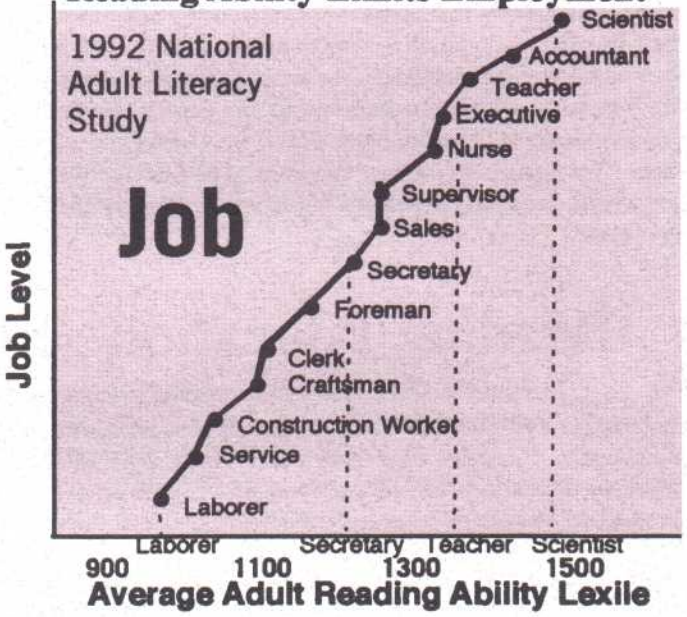
Job

Twenty-five thousand adults reported their jobs to the 1992 National Adult Literacy Study (Campbell et al., 1992). Their reading ability was also measured. **Figure 1** summarizes the relationship between reading ability and employment. In 1992, the average laborer read at 1000 Lexiles; the average secretary at 1200; the average teacher at 1400; the average scientist at 1500.

When we can see so easily how much increasing our reading ability can improve our lives, we cannot help but be motivated to improve, especially when what we must do is so obvious. If we want to be a teacher at 1400 Lexiles but read at only 1000, it is clear that we have 400 Lexiles to grow to reach our goal. If we are serious about teaching, the Lexile Framework shows us exactly what to do. As soon as we can take 1400-Lexile books off the shelf and read them easily, we know we can read well enough to be a teacher. But if we find that we are still at 1000 Lexiles, then we cannot avoid the fact that we are not ready to qualify for teaching, not yet, not until we teach ourselves how to read more difficult text.

Figure 1

Reading Ability Limits Employment

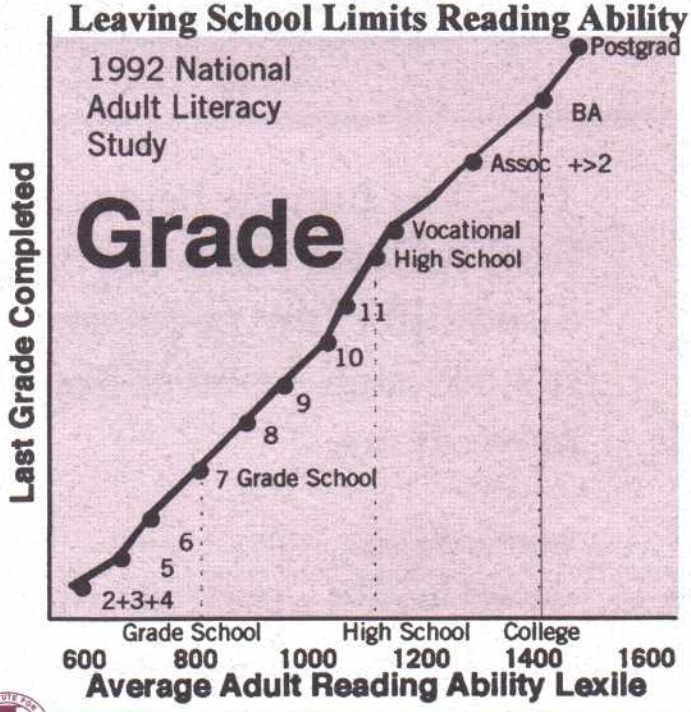


School

Reading is learned in school. 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. The average 7th grade graduate reads at 800 Lexiles. The average high school graduate reads at 1150 Lexiles. College graduates can reach 1400 Lexiles. For many of us, the last grade of school we successfully complete defines our reading ability for the rest of our lives. Once we leave school — and we no longer benefit from the reading challenges that school provides — we tend to stop learning. The overwhelming implication of **Figure 2** is that, if we aspire to become a truly literate society, then we must maintain schooling for everyone and help everyone stay in school as long as possible.

Figure 2

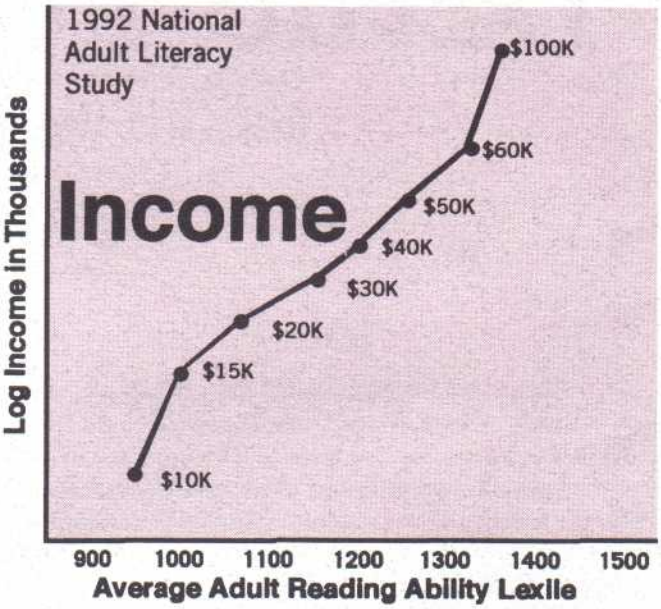
Leaving School Limits Reading Ability



Income

Reading ability also limits how much we can expect to earn. **Figure 3** shows the average incomes of readers in the 1992 National Adult Literacy Study at various Lexile reading abilities. From 1000 to 1300 Lexiles, each reading ability increase of 150 Lexiles doubles our earning expectations. If we read at 1000 Lexiles and want to double our potential, then we have to improve our reading 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 simple relationship in **Figure 3** makes the road to riches obvious and explicit. No need to berate students, "Do your home work!" Instead, we can show them, "You want more money? You want to be a doctor? Here is the road. Learn to read better. It's up to you. But we'll help you learn."

Figure 3
Reading Ability Limits Income



Reading Education

Education can only succeed if we connect learning to each learner's selfish motives. We need to involve our students individually, to engage their desires and arouse their drives. When we do that, student education will drive itself. Then, all we need do is to add support and guidance. Otherwise, we will continue to deceive ourselves into running a penitentiary system that keeps some troublesome kids off the street, but only for a while.

Remember, when we know text readability, all we need do to learn how well a student reads is to ask them to read a page or two aloud. If they succeed, we can give them a harder 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 high they can read.

Sources

Campbell, A., Kirsch, I.S., & Kolstad, A. Assessing Literacy: The Framework for the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics, U.S. Department of Education, 1992.

The authors are grateful to Ed Bouchard for helping with this report.

The way human beings learn best is by "discovering" first, because that's the only way to create the cognitive disequilibrium necessary for learning to take place. It's through re-inventing the wheel that the students move along.

Martin Brooks
quoted in "The Open Book", April 1992, p.2.

Using Lexiles

Benjamin D. Wright and A. Jackson Stenner

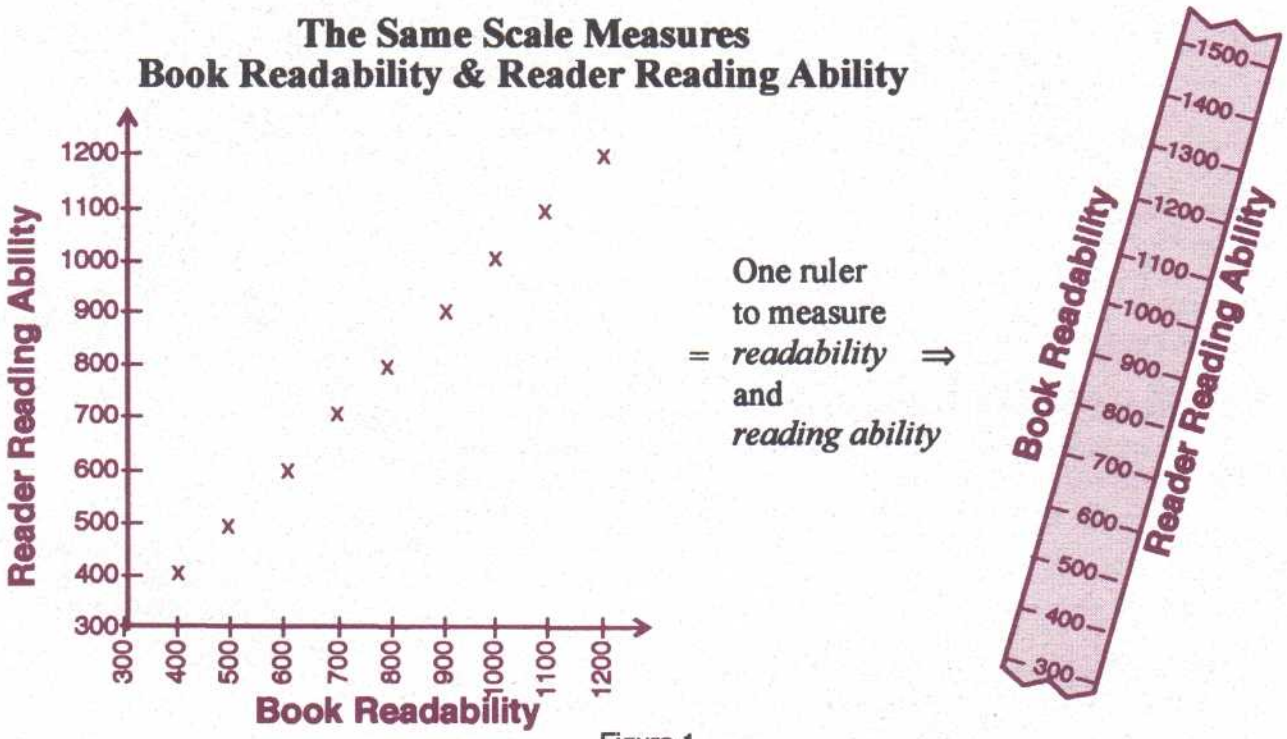


Figure 1

Books are brought into the Lexile Framework by Lexiling the books. Tests are brought into the Framework by Lexiling their items and using these Lexile calibrations as the basis for estimating readers' reading ability.

To write a Lexile test item, we can use any natural piece of text. If we wish to write an item at 1000 Lexiles, we select books that contain passages at that level. We select a 1000-Lexile passage and add a relevant continuation sentence at the end with a crucial word missing. This is the "response illustration." Then we compose four one-word completions, all of which fit the sentence but only one of which makes sense. Thus, the only technical problem is to make sure all choices complete a perfectly good sentence, but that only one choice fits the passage. The correct answer for the response illustration in **Figure 2** is "Use **repetition** for emphasis."

The aim of a Lexile item is to find out whether the student can read the passage well enough to complete the response illustrated sentence with the word that fits the passage. Lexiled items like this are available at the Lexile web-site, www.lexile.com. Anyone can use them, any time.

The Lexile Slice is a simple easy-to-write item type. But in practice, we may not even need the slice to determine how well a person reads. Instead, we may proceed as we do when we take a child's temperature. Since the Lexile Framework provides a ruler that measures readers and books on the same scale, we can estimate any person's reading ability by learning the Lexile level of the books they enjoy.

Figure 2

A 1000 Lexile Slice Test Item

You don't just establish a character once and let it go at that. Dominant impression, dominant attitude, dominant goal, all the rest — they must be brought forward over and over again; hammered home in scene after scene, so that the audience has no opportunity to forget them. Use _____ for emphasis.

A. humor
B. lighting
C. repetition
D. volume

READING RULER

The One-Minute Self-Report

When our child says "I feel hot!" we infer they have a fever. When a person says "I like these books," and we know the books' Lexile levels, we can infer that the person reads at least that well.

The Three-Minute Observation

To find out more about our child, we feel their forehead. The three-minute way to measure a person's reading is to pick a book with a known Lexile level and ask the person to "Read me a page." If they read without hesitation, we know they read at least that well. If they stumble, we pick an easier book. With two or three choices, we can locate the Lexile level at which the person is competent, just by having them read a few pages out loud. With a workbook of Lexile calibrated passages, we can implement the three-minute observation this simply by opening the workbook and turning the pages to give them succesive passages to read.

The Fifteen-Minute Measurement

To find out more, we use a thermometer to take our child's temperature, perhaps several times. For reading, we give the person some Lexiled passages ended with an incomplete sentence. To measure their reading ability, we find the level of Lexiled passages at which that person correctly recognizes what words are needed to replace the missing words 75 percent of the time.

The Lexile reading ruler connects reading, writing, speaking, listening with books, manuals, memos, and instructions. This stable network of reproducible connections empowers a world of opportunities of the kind that the inch makes available to scientists, architects, carpenters, and tailors.

Figure 3
Taking a Measure

Method	Temperature	Reading
One Minute Self-Report	I have a fever!	I like this book!
Three Minute Observation	You feel hot!	Read this page.
Fifteen Minute Measurement	Your temperature is...	Your Lexile is...

In school, we can measure which teaching method works best and manage our reading curricula more efficiently and easily. In business, we can Lexile job materials and use the results to make sure that job and employee match. When a candidate applies for a position, we can know ahead of time what level of reading ability is needed for the job and evaluate the applicant's reading ability by finding out what books they are reading and asking them to read a few sentences of job text out loud. This quick evaluation of an applicant's reading ability will show us whether the applicant is up to the job. When an applicant is not ready, we can counsel them, "You read at 800 Lexiles. The job you want requires 1000 Lexiles. To succeed at the job you want, you need to improve your reading 200 Lexiles. When you get your reading ability up to 1000, come back so that we can reconsider your application."

*The authors are grateful to Ed Bouchard for helping with this report.

Rasch At Work

Betty A. Bergstrom, Ph.D.
John A. Stahl, Ph.D.

Job task analysis (often referred to as practice analysis, audit of practice, task analysis, or role delineation study) is used to validate examinations by providing a link between performance on the job and examination content. Performing a job task analysis (JTA) helps ensure that examination content specifications are current and relevant.

A nursing subspecialty group undertook a job task analysis, with the ultimate goal of updating their certification examination. The data consisted of responses from 427 individuals who participated in a task analysis survey. The respondents were asked to rate a variety of tasks based on how frequently they were performed and how critical they were to professional practice.

The data were analyzed with the Rasch model, which positioned the tasks on two linear scales. Tasks were ordered based on their relative frequency or criticality.

Fig. 1

Frequency vs. Criticality Calibrations

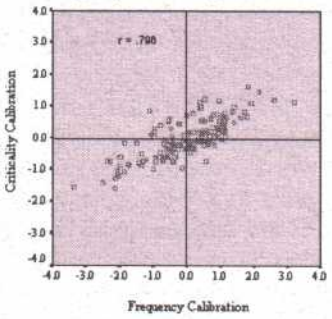


Figure 1 shows the frequency scale plotted against the criticality scale. Those tasks in the upper right quadrant are frequently seen and considered critical to practice by the respondents. Those tasks in the lower left quadrant are infrequently seen and not deemed critical to practice.

The frequency variable spanned a range of approximately 4 logits, with negative calibrations representing interventions rarely practiced and positive calibrations representing tasks frequently practiced. The criticality variable spans approximately 4 logits, with negative calibrations representing unimportant tasks and positive calibrations representing very important tasks.

Subject-matter experts reviewed these results. On the frequency and criticality scales, tasks having calibrations that fall below -.50 on either or both scales are potential choices for elimination from the examination content. These are tasks that are infrequently encountered, unimportant, or both. Subject-matter experts also reviewed items identified as misfitting.

Calibrations were transformed to relative percentage of questions on the examination using a procedure developed by Lunz, Stahl, and James (1989). This ensures that the tasks with the highest calibrations on the variable receive the highest relative percentage of items on the test and the tasks with the lowest calibrations receive the lowest percentage of items on the test.

By performing a job task analysis, an organization is able to determine the tasks and procedures most relevant to professional practice, and construct certification and licensure examinations that reflect these responsibilities. Using the Rasch model for calibration ensures that the tasks are on an equal interval scale. Plotting the criticality scale against the frequency scale allows subject matter experts to determine what tasks are both critical and frequently seen, thus enabling them to make informed decisions about content specifications. Test content can be balanced to ensure that tasks having the highest calibration on the latent variable receive the highest relative percentage of items on the test and the tasks having the lowest calibration receive the lowest percentage of items on the test.

- edited by Andria Brown

Bibliography

Lunz, Mary E.; Stahl, John A.; James, Karen (1989). Content validity revisited: Transforming job analysis data into test specifications. Evaluation and the Health Professions, 12 (2). 192-206.

BETTY A. BERGSTROM, Ph.D. Vice President of Program Management and Psychometric Services. Dr. Bergstrom directs Research and Psychometric Services at Computer Adaptive Technologies, Inc. She is responsible for all research, psychometric, and custom software clients at CAT, Inc. Dr. Bergstrom consults in the areas of Item Response Theory (IRT), equating, standard setting, item banking, computerized testing, and adaptive testing.

JOHN A. STAHL, PH.D. Research Scientist. Dr. Stahl is a research scientist at CAT, Inc. where he oversees Responsibilities theoretical and practical applications in item response theory for use in computerized adaptive tests and judge-mediated assessments. Dr. Stahl oversees the technical development of the CATBuilder and the CATAdministrator. He also teaches many of the sessions in the "Future of Computerized Testing is Now!" seminar series.

Testing Smarter with Technology?

Anne Wendt, Ph.D., R.N.
National Council of State Boards of Nursing, Inc.
Chicago, IL

Abstract

This article briefly discusses the factors affecting the National Council of State Boards of Nursing's decisions to use computerized adaptive testing to administer the National Nursing Licensure Examination (NCLEX(r) examination). Practical issues associated with implementing computerized adaptive testing for large-scale, high-stakes test such as the National Council's Licensure Examination (NCLEX(r)) are discussed.

Have you wondered why so many tests are now being administered via computer? Using the National Council of State Boards of Nursing licensing examination (the NCLEX examination) as a model, we will examine some of the issues surrounding computer based testing, particularly the advantages and potential problems that can occur with its implementation.

Reasons for Using Technology in Testing

Aside from the excitement of using up-to-date technology for testing, one of its major advantages is convenience to the candidates. Prior to April 1994, the NCLEX examinations were paper-and-pencil examinations which were administered to about 190,000 nursing candidates per year. Each examination was administered twice a year for one day (NCLEX-PN exam) or two days (NCLEX-RN exam). Now candidates call and schedule their examination at a time and place that is convenient for them. Candidates are tested via computer in one of approximately 250 Sylvan Technology Centers that are located in each state or U.S. territory. In addition, candidates are tested in a quiet, private, self-paced setting rather than in the large auditoriums or halls when everyone began and ended their examination at the same time.

Another major advantage for the National Council of using technology for testing is that a large volume of candidates can now be tested almost daily on a year-round basis. In addition, candidate results are reported to the respective Boards of Nursing within 48 hours. Thus, candidates who are competent are able to join the workforce much more quickly. And Boards of Nursing are able to quickly identify those individuals who are not competent to practice (Zara, 1996). Therefore, by taking advantage of computer technology, Boards of Nursing are better able to perform their role of public protection while licensing nurses in a timely manner.

A third advantage to the National Council of utilizing advances in technology is to improve the security of the examination. Major causes of security breaches for paper-and-pencil examinations are lost and/or stolen test booklets, copying answers from other test-takers during the examination, and hiring impersonators to take the examination (Scheuneman, 1997). The use of the computer adaptive testing has virtually eliminated the first two causes of security breaches. With computerized testing, examination materials can be encrypted. The adaptive nature of computerized adaptive testing (CAT), where each candidate is administered a unique set of test questions tailored to his/her competence level, eliminates the possibility of copying answers from the person taking the examination in the next testing cubicle. Technology is also used to enhance security measures. For example, the National Council uses strict security procedures which include a sign-in log, appropriate picture identification, a digital photograph taken at the site which can be forwarded with the examination results, thumbprinting, and video monitoring of the testing event. All of these measures help to ensure that the correct candidate takes the examination and that the testing event (as much as possible) is the same for all candidates.

Lastly, use of CAT for the NCLEX examination allowed a reduction in testing time without a loss in the precision or accuracy of the results. That is because candidates do not "waste time" trying to answer questions that are too hard or too easy for them. Basically, that is the goal of CAT— to determine competency based on the difficulty of the questions answered correctly, rather than the number of questions which are answered correctly as is the case with many paper-and-pencil or computer linear examinations (Wainer, 1990; National Council, 1995). While every candidate's examination meets the NCLEX Test Plan, only questions which contribute to the measurement of the individual candidate need to be administered. For the NCLEX-RN examination, the testing time decreased to five hours from two days. And, because NCLEX CAT is a variable-length examination, many of the candidates taking the examination finish in less than the allotted 5 hours. Thus, the increased efficiency of computerized adaptive testing for the NCLEX examination can be translated into savings to the candidate and organization.

In summary, there are a number of compelling reasons to take advantage of changes in technology to begin "testing smarter." By implementing computerized adaptive testing, the National Council has been able to increase test security, shorten testing time, enhance Member Boards' mission of public protection, and streamline the testing process.

Practical Issues

While the decision to take advantage of technological changes may seem to be an easy one, there are some practical issues that need to be addressed. One of the major ones is how to maintain the psychometric soundness and legal defensibility of a CAT administered examination. The National Council spent a great deal of time and effort to ensure that an NCLEX CAT examination would be legally defensible. Much

of this process is documented in the Collected Works on the Legal Aspects of Computerized Adaptive Testing (National Council, 1991). In addition, the National Council conducted field tests and a large-scale beta test to ensure the psychometric soundness of the NCLEX examination administered by CAT methodology (Way, 1994). The preparation for implementing CAT for the NCLEX examination took considerable time and effort. Do not underestimate the time, effort, and money that it will take to address these issues.

Now let us move on to the more "practical" aspects of testing smarter with technology. Two of the major issues that need to be addressed are the cost of setting up a system for computer delivery of the multiple-choice test questions and the quantity of questions that are available for converting to computer delivery. The former may be easily achieved if the questions are already formatted and available in electronic files. The latter may be quite costly. There are several reasons that many questions (items) are needed for CAT. First, large numbers of items are needed to ensure that there is an "acceptably low" level of item overlap between candidates. Limiting item exposure in this way prevents the items from becoming common knowledge among candidates. In addition, items are needed in all areas of a test plan at all difficulty levels to ensure precise estimates of candidates with varying competence levels. Because the field of measurement does not have a good way to perform "on-line item calibrations", the items will need to be tried out (given as unscored items to candidates in order to gather statistical information about the item) before administering the items as "operational" scored items. This extensive task is very time-consuming.

It should also be noted that computerized adaptive testing items may need to be classified more carefully using dimensions other than the test plan categories. When a paper-and-pencil test form is produced, a content expert selects the items and reviews the composition of the test form for such things as cueing and overlap. This is not possible when the computer automatically selects each candidate's items based on an algorithm. Therefore, items may need to be classified for additional dimensions. Additional and in-depth item classification is especially important in a health-care related examination where there are likely to be frequent changes in practice and/or terminology necessitating frequent reviews of the items in the pool.

More detailed item classification is also important when considering item pool needs. With large item pools, there may be no easy or effective way to determine voids in content areas without more detailed item coding. With smaller item pools it may be possible for one or two people to be the "gurus" of the item pool, i.e., know what specific concepts are and are not addressed in the items. But with any more than about 3,000 items, in-depth knowledge of the pools is impossible. Detailed item classification allows for a more in-depth knowledge of test content and an easier determination of voids in that content.

Now let us turn to some administration issues. As noted previously, one of the compelling reasons for computer-administered examinations is the reduction in security issues. However, it should be noted that continued vigilance in matters of security is absolutely necessary. Candidates for high stakes examination may try to find ways to circumvent the security measures. Any irregularities need to be investigated. The time and effort that this may take should not be underestimated. Candidates may complain about adverse environmental conditions such as excessive heat or noise. These complaints also need to be investigated and followed up. With multi-site, daily testing it is important to be diligent in tracking issues and trends. To effectively handle all of these administration issues, additional staffing may be required.

These practical issues are not intended to be an exhaustive list of what can be encountered when taking advantage of technological advances. It is up to the test-sponsoring organization to assess if the advantages to its stakeholders outweigh the disadvantages. In particular, some of the issues to consider are "Will computer-based testing save time or money for the program or candidates?" "Will security improve?" "Will better decisions about candidate competence be made?"

References

National Council. (1991). Collected Works on the Legal Aspects of Computerized Adaptive Testing. Chicago: National Council of State Boards of Nursing.

National Council. (1995). The NCLEX Process. Chicago: National Council of State Boards of Nursing.

Scheuneman, J. (1997). Testing and measurement issues: potholes on the road to computer-based testing. CLEAR Exam Review. Volume VIII, Number 1. Kentucky: Council of Licensure, Enforcement, and Regulation.

Wainer, H. (1990). Computerized adaptive testing: a primer. Hillsdale New Jersey: Erlbaum Associates, Publishing.

Way, W. (1994). Psychometric results of the NCLEX beta test. Paper presented at the Annual Meeting of the American Educational Research Association.

Zara, A. (1996). "Overview of a successful conversion." In Computer-based examinations for board certification. Evanston, Illinois: American Board of Medical Specialities.

Anne Wendt, Ph.D., R.N.

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.

She is the author and co-author of numerous articles and has presented papers nationally. She continues to be actively involved in various research projects, particularly in the areas of item development and test construction.

Adjusting for Rater Severity Over Time

Thomas R. O'Neill

Performance assessments are often thought to have greater validity than multiple-choice tests because the rated behavior more closely approximates the behavioral domain of interest than does merely asking questions about it. For example, expert judges rating a student on a karate test want to know if the student knows how to strike with penetrating force. It seems obvious that asking the student to break a brick is more informative than asking the student to answer questions about breaking a brick. While breaking a brick is usually unambiguous with regard to success, other activities such as judging technique must be graded by raters. Because raters often have different individual standards of excellence, the reproducibility of estimates derived from ratings is sometimes questioned. Any given rating will be influenced not only by the examinee's ability and the task's difficulty, but also by a third facet, rater severity.

In order for measurements to be meaningful, differences in raters must be accounted for, so that all results are expressed from the same frame of reference. The extension of the Rasch (1960/1980) model to the Many Facet Rasch Model (MFRM, Linacre, 1989) has made accounting for rater severity possible by placing rater severity in the same frame of reference as item difficulty and examinee ability. The MFRM estimates each rater's severity, each project's difficulty, and/or other such facets, and removes their influence before computing an examinee's ability. An examinee's measure is independent of which rater graded them and which tasks they performed.
An alleged drawback is that with each additional link required to connect a test form back to its original scale, more error accumulates. However, it is often overlooked that with each successive administration, more historical data is available to guide the test development process.

A linking strategy is usually employed to align the scale defined by the current test administration with the original scale, thus the same scale is maintained across several administrations. In multiple-choice test, this linking is usually accomplished using several items common to both the current test form and the preexisting scale. Once the difficulty of the items on the new are aligned with the preexisting scale, an examinee taking two forms of the test will receive a comparable measure even when the test forms are different in difficulty.

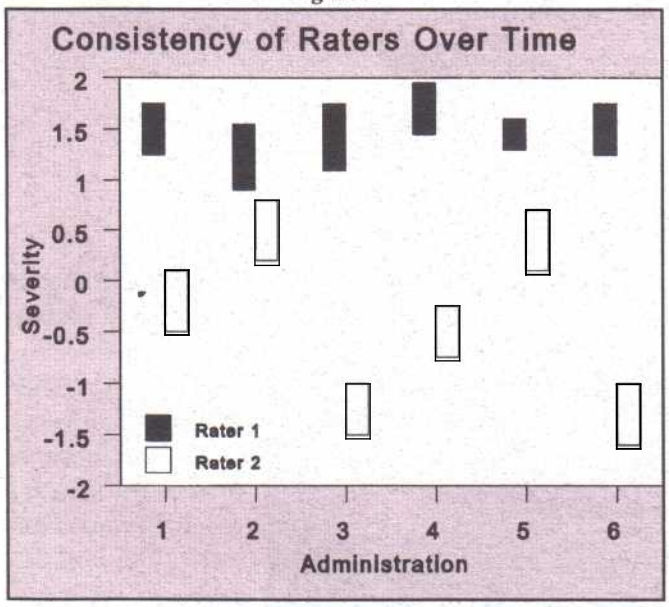
In performance assessments, the difficulty of the prompts from the current form must be aligned with the prompt's preexisting scale, but this alone is not enough to make the examinee measures comparable. The severity of the raters must also be aligned. It is important that the severity of new raters is expressed in the same frame of reference as that of the original raters. Using common raters to link together two test administrations requires that the common raters maintain a uniform degree of severity across administrations. However, actual raters occasionally violate this requirement and thereby potentially thwart our intention to carry forward the same scale. For this reason, it is important to use historical data to identify those raters who are most likely to maintain a uniform degree of severity across administrations.

As part of the equating process, rater stability is verified from administration to administration. This is done by comparing the severity of several common "anchor" raters on the current administration with their degree of severity from the prior administration, and then checking that their severity on the current administration places them in the same relative position as in the past. When their relative positions hold, it is reasonable to conclude that their severity has not changed. In cases where only one or two of the anchor raters have changed positions, it is reasonable to conclude that those one or two raters have changed their degree of severity and should be treated as new raters, but the rest of the anchor raters can be used to link the new raters to the established scale. But when several raters change places and the number of anchor raters is few, it becomes more complicated to determine which of the anchor raters changed their severity and which remained the same.

To prevent this from happening, psychometricians try to employ as many stable pre-calibrated raters as possible, so that any anomalous raters will stand out more clearly. While it can never be known in advance exactly how severe a particular rater will be on any given occasion, a rater's past performance can suggest how severe they will be in the future. Thus, historical information can be helpful to psychometricians who are organizing or equating performance assessments across administrations. By plotting a rater's severity with their error bands (±2 SEs) across administrations (**Figure 1**), psychometricians can verify that things are going well or identify problem areas. A method to do this can be found in Objective Measurement: Theory into Practice (volume 5).

Analyzing rater severity overtime should be part of the ongoing equating procedure because it can aid developers with historical data in making decisions about raters. For example, a psychometrician may select a few raters to participate in several consecutive administrations for the purpose of maintaining the same frame of reference for rater severity. Common raters should be selected on the basis of their documented ability to maintain a uniform level of severity. Armed with historical information **(Figure 1)**, psychometricians can seek out

Figure 1

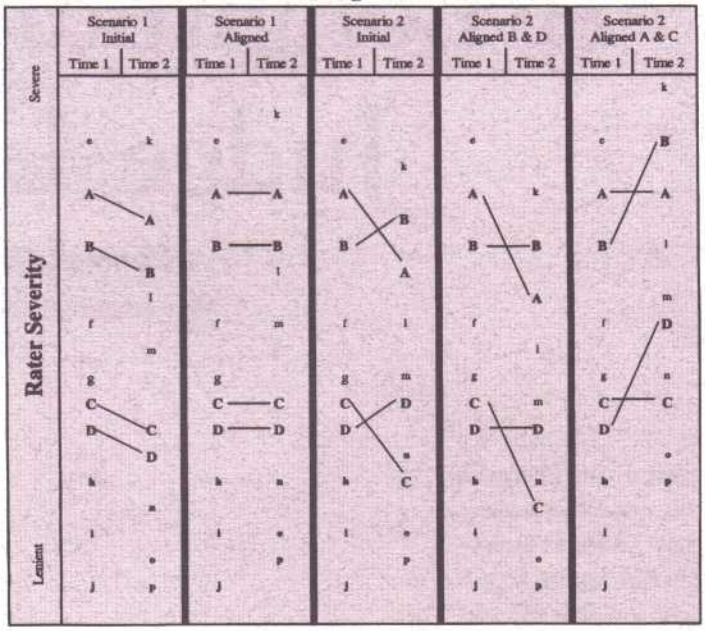


stable raters like Rater 1 for this purpose. Others, like Rater 2, can still be used across administrations because their degree of severity is consistent within administrations, but knowing their across-administration degree of severity has more variance, the psychometrician would not want to use them as a link back to the initial scale. Rater 2 should be thought of as a new rater each time he grades.

Viewing rater severity in this manner can generate hypotheses regarding how individual raters behave over time. When the psychometrician thinks that there has been a shift in a rater's severity and that the new level severity is likely to be stable, the psychometrician should update the rater calibration bank with the new severity calibration.

The most obvious information noticeable from **Figure 1** is which raters are consistent and which are erratic across administrations. This information can be used to select anchor raters, but it can also be used after the data has been collected. Suppose that out of ten raters, only four raters, A, B, C, and D had a known degree of severity **(Figure 2)** established from earlier administrations. Ideally, one would hope for results similar to the second administration as found in scenario 1 (initial). Because the four raters maintained their relative position from each other, aligning the common raters is a simple matter (scenario 1, aligned) which allows the severity of raters K through

Figure 2



P to be expressed in the same frame of reference as raters A through J.

However, suppose that two of the common raters changed their severity by approximately the same amount on the second administration as represented in scenario 2 (initial). How would the psychometrician know if A and C became more lenient (scenario 2, B & D aligned) or if B and D became more severe (scenario 2, A & C aligned)? Either scenario seems equally plausible. A potential answer is to review the historical performance of the four raters. It seems probable that the historically more stable raters would be less likely to be the ones who changed.

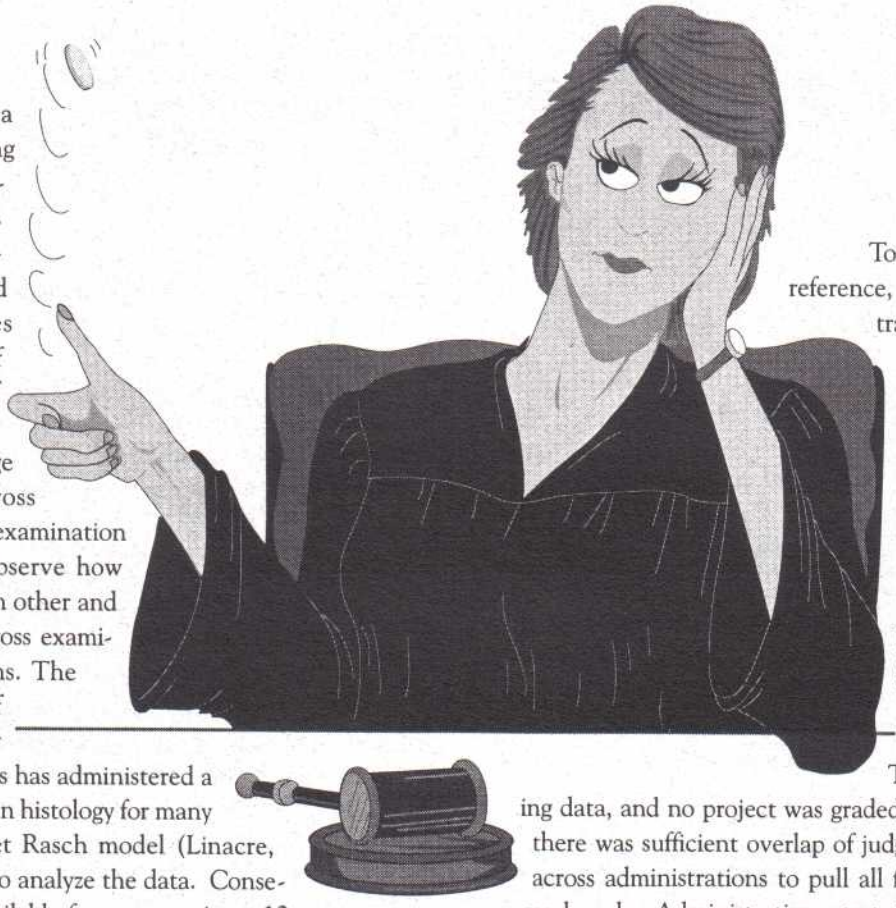
To prevent the above scenario, enough common raters should be employed so that if a small percentage of raters change in severity, it will be easy to identify which raters changed. Reviewing the historical data can allow the psychometrician to make a good guess that, given the available pool of stable, pre-calibrated raters, (1) which raters should be selected, (2) how many of the raters are expected to change severity during this administration, and (3) how many raters will be needed to clearly identify those who have changed severity.

Thomas R. O'Neill is Manager of Research & Analysis for the American Society of Clinical Pathologists. His professional interests include performance assessment, Computerized Adaptive Testing (CAT), and the promotion and creation of understandable and useable measurement. His personal interests include Korean Tang Soo Do (karate), zymurgy (beer brewing), and travel.

email: tom_o'neill@ascp.org (the apostrophe is important!)

A Longitudinal Study of Judge Leniency

Mary E. Lunz, Ph.D.
Measurement Research Associates, Inc.

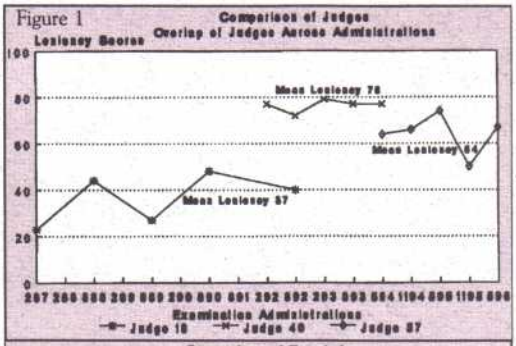


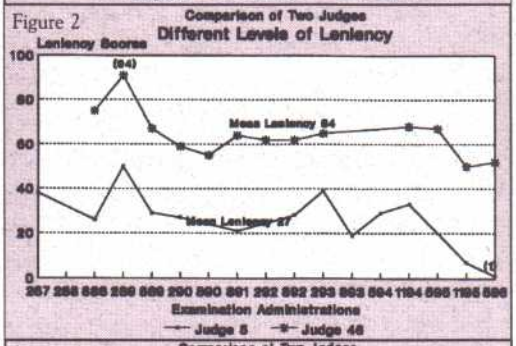
The judge is a critical part of scoring any clinical examination. However, little is known about long-term consistency and leniency. Do judges change their level of leniency over time; if so, in what directions? This study tracks judge grading patterns across ten years of clinical examination administrations to observe how judges differ from each other and within themselves across examination administrations. The Board of Registry of the American Society of Clinical Pathologists has administered a clinical examinations in histology for many years. The multi-facet Rasch model (Linacre, 1989) has been used to analyze the data. Consequently, data were available for constructing a 10-year longitudinal study of judge performance. The clinical examination has four facets: 1) candidates, 2) judges, 3) projects, and 4) tasks. Over the ten years there were 4,683 candidates, 57 judges, and 53 projects. Three tasks were graded at each administration. Two were graded as 1=acceptable and 0=unacceptable and the third task was graded on a four-point scale as 3=excellent, 2=acceptable, 1=marginal, 0=unsatisfactory. The same grading scales were used for all administrations. Candidate performances were randomly assigned to judges. Each candidate was judged on the three tasks for 15 projects, with input from three judges. All judges graded examples of all projects during each administration.

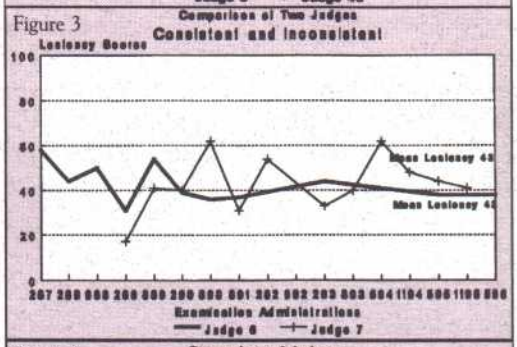
To construct a frame of reference, data from 17 administrations were pooled and analyzed together. This placed all examination administrations for ten years on the same "benchmark" scale. The FACETS program (Linacre, 1994) was used to calibrate candidate ability, judge leniency, project difficulty, and task difficulty on this scale. There was a lot of missing data, and no project was graded more than once. But there was sufficient overlap of judges, projects, and tasks across administrations to pull all facets onto the benchmark scale. Administrations started in February, 1987 (labeled 287) and continued semi-annually through May, 1996 (596).

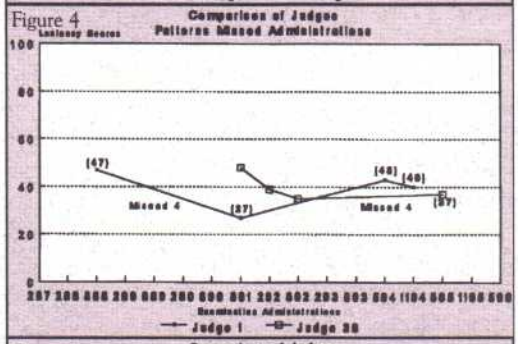
After the benchmark scale was constructed, individual examination administrations were re-analyzed separately. The difficulty estimates for the projects and the tasks, as well as the candidate ability measures from the benchmark scale, were used to anchor the individual examination administrations. The non-anchored facet across administrations was judge leniency. This enabled differences in judge leniency to be tracked across administrations. The multi-facet judge leniency estimates were transcribed to scaled scores so that 0 points marked the most lenient judge and 100 marked the most severe judge.

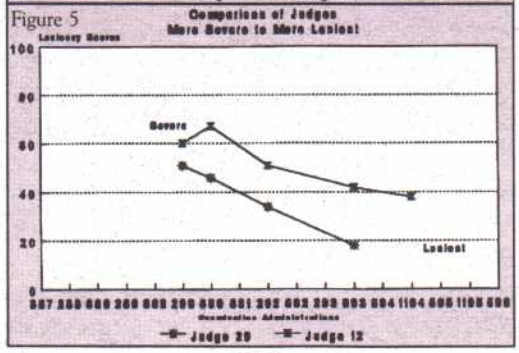
On average, judges graded in six administrations, but the range was 1-15 administrations. Different subsets of judges graded during each administration. However, there were always some judges that overlapped among administrations. **Figure 1** shows that judges 18 and 57 were linked with judge 40. Mean candidate ability estimates across administrations were verified as not significantly different among test administrations. Most judges graded in some administrations and skipped others. Some judges graded many sessions, while others graded few. Some judges varied among administrations, while others were extremely consistent. The graphs show examples of judge grading patterns across administrations. **Figure 2** shows the comparison of a relatively severe and a relatively lenient judge. The mean leniency of judge 46 was a scaled score of 64 points, while the mean leniency of judge 5 was a scaled score of 27 points. Each of these judges graded in 13 administrations and varied within 20 points of their average leniency across all examination administrations. **Figure 3** shows judges who are consistent and inconsistent in their leniency among administrations. Each of these judges graded at 10 of the 17 administrations. The average leniency of both of these judges was a scaled score of 43; however, judge 7 tended to vary in overall leniency at each administration, while judge 6 showed little variance after the first several examination administrations, even when administrations were missed. **Figure 4** shows that judges are consistent in their leniency even when they do not grade in consecutive examination administrations. Judge 38 graded three consecutive administrations, then missed four consecutive examination administrations, but stayed within a 10-point leniency range. Judge 1 graded in one administration, then missed four administrations, then graded one administration, then missed four administrations, but remained within a 10-point leniency range. **Figure 5** shows two judges who moved from relatively severe to relatively lenient. Some sessions were missed, but the pattern of becoming more lenient is obvious for these judges. The study shows that clinical examination data from different examination administrations can be placed on a benchmark scale when there are commonalties that link examination administrations using the multi-facet model. Some judges were consistent across years; however, some were less consistent, possibly because of limited grading experience, educational or personal changes, or technical experience.











Mary E. Lunz, Ph.D.

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., which provides psychometric services to medical specialty, dental specialty, and allied health certification organizations. Her specialty is the development and analysis of performance examinations, especially oral interview examinations. (MeasResInc@aol.com)

PUTTING THE PSYCH IN PSYCHOMETRICS

Larry H. Ludlow, Ph.D.
Boston College

Imagine that you have just spent an hour explaining the operation of the Newton-Raphson iteration technique to your second-year doctoral students. You are standing in front of the class and feeling quite pleased with yourself. The board and screen are covered with equations and graphs. One palm is damp from continuous use of the laser pointer, the other is chalk-covered. You are slightly out of breath, yet strangely energized.

You have lectured on this topic a half dozen times in your career. Tonight, however, you feel you have actually "taught" the students how the technique works and why knowing about it is important. You even believe you have made the topic interesting, if not exactly exciting. Bottom-line, you have answered the ultimate questions, "So what?" and "Who cares?"

As you look at the students, awaiting their applause, a thought occurs, "What are they thinking?" You ponder this question as they file out of the room. There is no applause, no wave, no cheers, not even a "Nice job, Doc!" You wonder how they describe this class to their friends. What visual images do they construct for their audiences?

Over the next couple of days you ask a few students what they thought about the lecture. Did they understand it reasonably well? Was it clear? Did it make sense? Was it at their level? Where were the tough parts? Where did they begin to lose it? Their responses are nondescript—it was fine, it was interesting to see how the parts fit together, it made sense at the time, it was challenging but OK. Their responses, while somewhat supportive of your efforts, don't leave you satisfied. So you decide to try something unusual in the next class.

At the start of the next class session you ask them to "draw a typical classroom experience that includes me, yourself, and everything else that represents that classroom experience." The class ripples with giggles. Students look at one another. Puzzled expressions are exchanged. Whispers and groans are heard. Some of them look at you as if you have gone really weird on them this time. Eventually they begin to draw.

When everyone has finished, you ask them to turn the paper over and write an explanation of the scene. In addition, you ask them to write what they think the drawings convey about the course that is not conveyed in the scannable course evaluation forms.

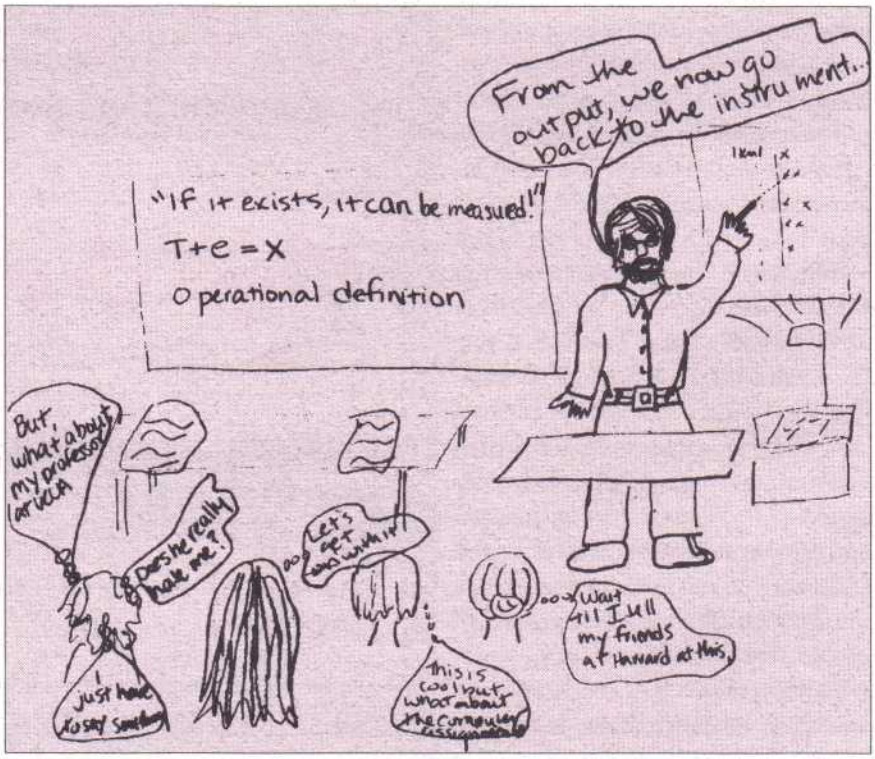


Figure 1: Doctoral student drawing of an intelligible presentation

Figures 1 and 2 were drawn by two students at the start of one of the final sessions of my spring 1998 psychometrics class. From my perspective, **Figure 1** reflects a relatively positive classroom experience. The "statement bubble" over my head makes sense, there is a discernible variable map on the overhead, there are correctly stated phrases on the blackboard, and the students are all engaged and awake at some level.

Figure 2, however, is disturbing. Although I see an interpretable diagram of category characteristic curves, there is absolutely no doubt that this is a scene conveying an environment of confusion.

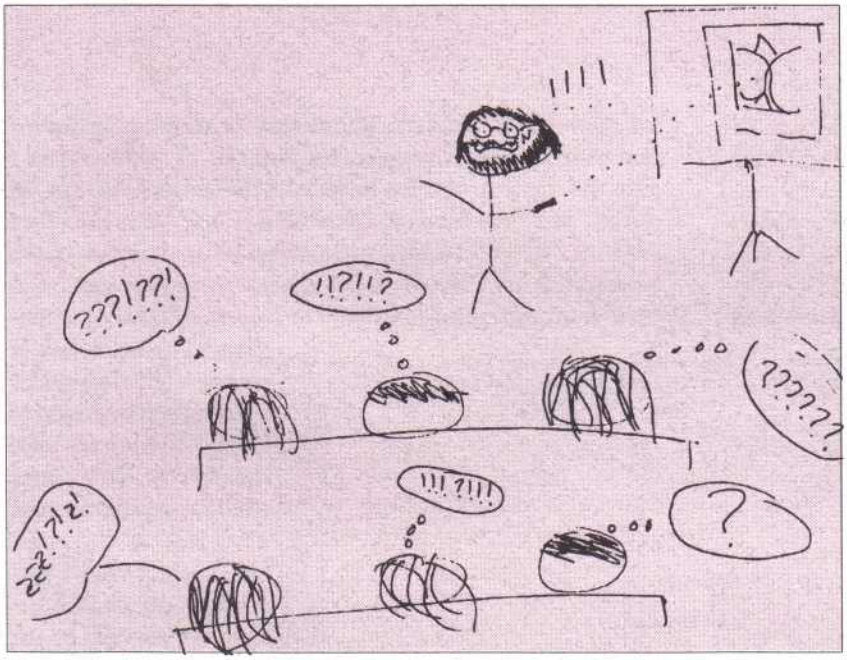


Figure 2: Doctoral student drawing of an unintelligible presentation

Since 1983 I have taught graduate level statistics and psychometrics courses. At the close of each semester I pass out the standard course evaluation forms required at my university. Over time, however, it has become apparent that the evaluations give me little information about student experiences in my classes. The forms are also very poor indicators of teacher effectiveness (Ludlow, 1996). When I learned in 1995 that elementary and middle school students were drawing interesting pictures of their classrooms that were useful to teachers (Haney, et al., 1998; Gulek, in progress), I decided to try the drawing technique in my classes. I now have drawing data from seven different graduate courses that I teach in measurement, evaluation, and statistical analysis—Interpreting & Evaluating Research, Statistics I, Statistics II, Multivariate I, Multivariate II, Psychometrics, and Seminar in Educational Research.

To my amazement and delight, the drawings are rich beyond any expectation I held for them. In fact, the gestalt portrayed in these drawings is so powerful that I am still unable to adequately explain the analytic process by which I try to understand their meaning. The problem, of course, is how to interpret and explain these drawings in some way that is not self-serving, idiosyncratic, or arbitrary.

To that end I am pursuing a variety of research questions. Basically, I am curious about how a viewer interprets and describes the information in drawings generated for course evaluation purposes. To a certain extent I am asking, "How do I get past the bean counting of feature detection analysis in order to expose unconscious expression and impression?" More practically, I am exploring: (a) what is important in these drawings? (b) what are students trying to say about a particular course and instructor? (c) what is unique and different about the courses? (d) which patterns are similar across courses? and (e) how can these drawings be systematically analyzed? Finally, how can qualitative drawing data be combined with quantitative course evaluation data to yield a richer understanding of the psychological dynamics underlying student evaluations of a course?

These questions are addressed in a number of articles in progress (titles subject to change). These include articles on self-inquiry and reflection on teaching practice, statistics education, alternative modes of evaluation of teaching in higher education, and the analysis of qualitative data by a non-qualitative researcher.

Sound interesting? If so, and if you think you have a relatively thick skin, then ask your students to draw you and themselves near the end of one of your next classes.

References

Gulek, C. (in progress). Using multiple means of inquiry to gain insight into classrooms: A multi-trait multi-method approach. Ph.D. dissertation: Boston College, Department of Educational Research, Measurement and Evaluation.

Haney, W., Russell, M., Gulek, C. & Fierros, E. (Jan/Feb. 1998). Drawings on education: Using student drawings to promote middle school improvement. Schools in the Middle: Theory into Practice, 38-43.

Ludlow, L.H. (1996). Instructor evaluation ratings: A longitudinal analysis. Journal of Personnel Evaluation in Education, 10, 83-92.

Ludlow: Psychometrics drawings: 5/7/98

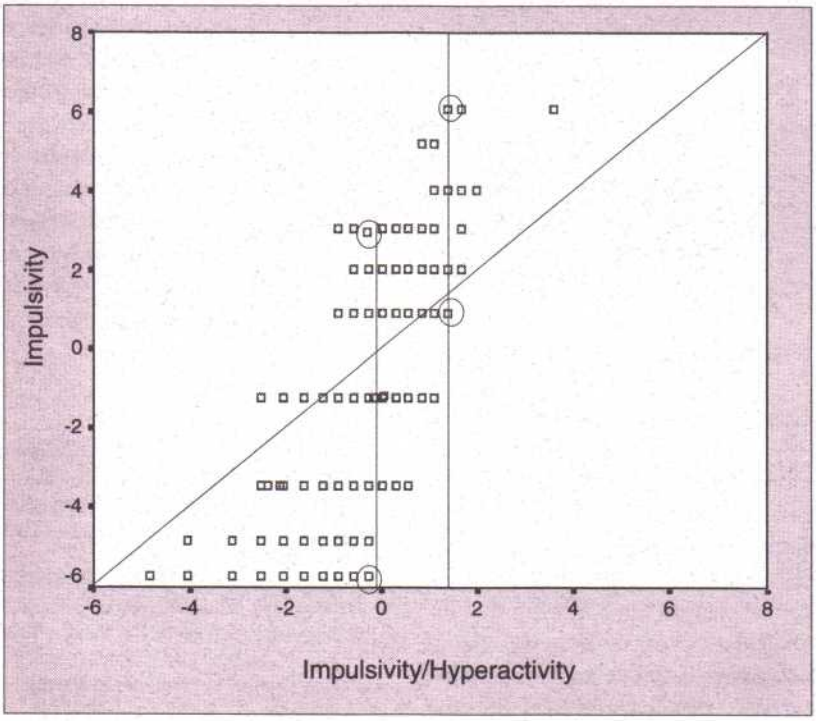
Attention Please!

Dimensions of Attention Deficit

Everett V. Smith, Jr., Ph.D.
The University of Illinois at Chicago

PSYCHOLOGY

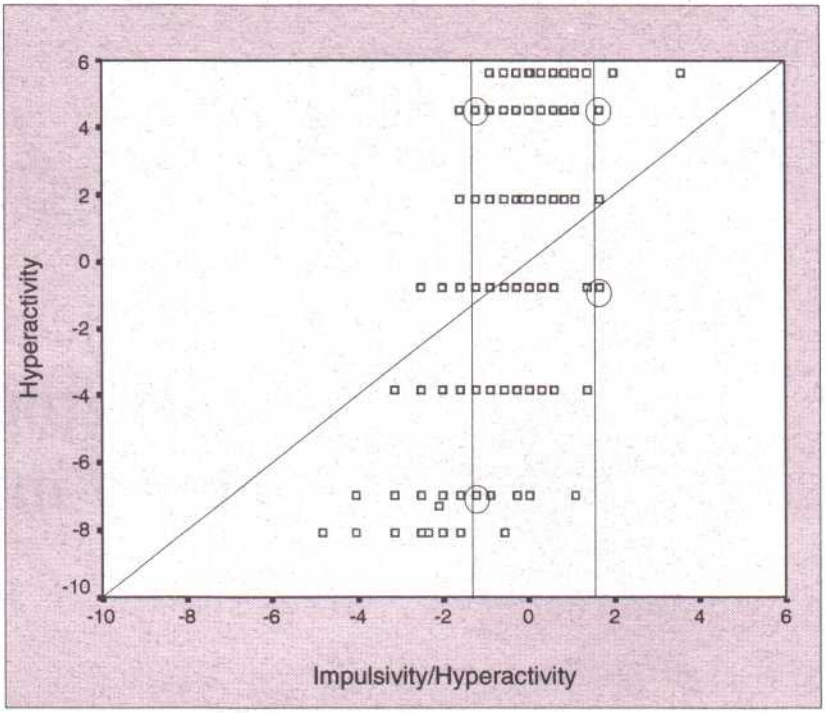
ADHD is characterized by inattentiveness (e.g., fail ing to finish schoolwork), impulsivity (e.g., interrupting conversations), and hyperactivity (e.g., always "on the go"). Investigations into the dimensionality of ADHD have stopped at gaining evidence to support the diagnostic subtypes advocated by the DSM-IV (i.e., Inattentive, Impulsive/Hyperactive, and combined Inattentive and Impulsive/Hyperactive) (APA, 1994). This paper demonstrates how to use features found in WINSTEPS to examine the Impulsivity/Hyperactivity dimension of ADHD and investigates the existence and potential utility of secondary dimensions. The data (n=317) used were obtained with the Adult Behavior Checklist - Revised (ABC-R), a screening assessment for ADHD in college students. The ABC-R addresses concerns noted by Smith and Johnson (1998) that several items on the original Adult Behavior Checklist contained more than one concept (e.g., work and school) and should be divided into two separate items.



The seven items representing Impulsivity/Hyperactivity (**Table 1**) fit the Rasch Rating Scale model. The correlations among standardized residuals were then analyzed into principal components. The purpose of this analysis was to investigate if order exists among the information remaining in the standardized residuals after accounting for the primary dimension of Impulsivity/Hyperactivity. Two sets of items emerged from this analysis. Set one included: "You act as if you are 'on the go'" and "You act as if 'driven by a motor.'" The behaviors in set two were: "You blurt out answers before questions have been completed," "You have difficulty awaiting your turn," and "You interrupt others (e.g., butt into conversations or activities)." The behaviors in set one seem to represent a secondary dimension of Hyperactivity; set two, Impulsivity. How useful are these two dimensions compared to the combined Impulsivity/Hyperactivity dimension already constructed by the Rasch analysis? By anchoring the items within each set at the calibrations found in the construction of the Impulsivity/Hyperactivity dimension, one is able to estimate measures for each of the secondary dimensions on the same scale that was constructed for the Impulsivity/Hyperactivity dimension. A simple bivariate plot of measures from the secondary dimensions and measures from the Impulsivity/Hyperactivity dimension will show whether measures from the secondary dimensions have clinical value.

Figure 1 plots measures from the Impulsivity/Hyperactivity dimension against measures from the Hyperactivity dimension. Examine for example, the participants with estimated

measures of -1.23 logits (n=24, each point represents more than one individual) on the primary dimension of Impulsivity/Hyperactivity. These participants would not be candidates for further evaluation because their measures are below the suggested cutoff of 1.60 logits. However, look at the range of their measures on the Hyperactivity dimension. The values range from -6.98 to 4.52 logits. Based on these secondary measures, several participants may need assistance with specific aspects of this disorder. Now examine the participants with estimated measures of 1.66 logits (n=4). These participants are candidates for further evaluation. Their corresponding measures on the Hyperactivity dimension range from -.77 to 4.52 logits. Participants with low measures on the Hyperactivity dimension may not need the more comprehensive evaluation to address these types of symptoms. Similar interpretations can be made from the information in **Figure 2**, which plots measures from the Impulsivity/Hyperactivity dimension against measures from the Impulsivity dimension. From the information in these Figures, it appears that interpretation of measures from the Impulsivity/Hyperactivity dimension without regard to measures from secondary dimensions may direct the evaluation and treatment of symptoms associated with ADHD.



It is hoped that this example will encourage researchers to 'dig' deeper in their investigations of dimensionality rather than terminating their investigations once support for a priori dimensions is obtained. First, the additional information found in the secondary dimension has the potential for enhancing a clinician's ability to evaluate symptoms and plan appropriate interventions. Second, the details provided by the secondary dimensions may assist with the interpretation of the theoretical constructs associated with the disorder under investigation.

Table 1

Items for the Impulsive/Hyperactive dimension

You have difficulty remaining quiet during leisure activities.

You act as if you are "on the go."

You act as if "driven by a motor."

You talk excessively.

You blurt out answers before questions have been completed.

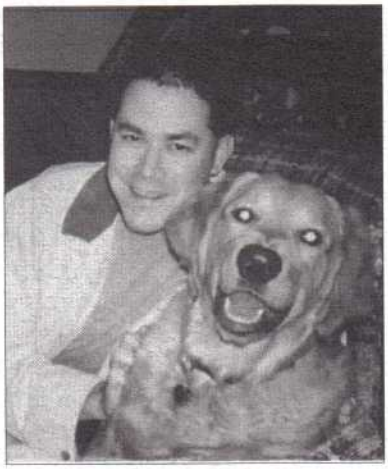
You have difficulty awaiting your turn.

You interrupt others (e.g., butt into conversations or activities).

References

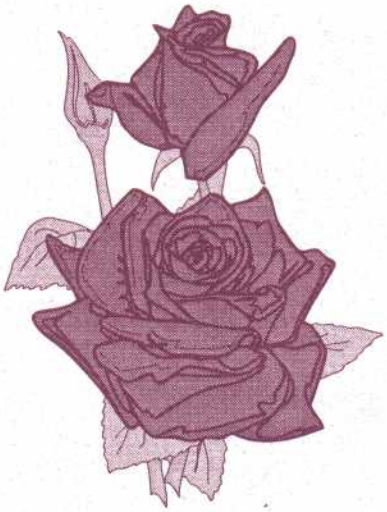
American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, D.C.: Author.

Smith, Jr., E.V., & Johnson, B.D. (1998). Factor Structure of the DSM-IV Criteria for College Students using the Adult Behavior Checklist. Measurement and Evaluation in Counseling and Development, 31, 164-183.



Everett Smith is an Assistant Professor of Educational Psychology at the University of Illinois, Chicago. He has recently developed a Ph.D. curriculum in Applied Measurement and Statistics that will begin accepting students in the Fall of 1999. The measurement aspect of the new curriculum will focus on objective measurement theory and applications. Current research interests include the development of a screening assessment for Attention Deficit Hyperactivity Disorder and the measurement of study skills self-efficacy in community college students. Other interests include biking, tennis, and spending time with Calvin (pictured) and Hobbes. Please e-mail me with any questions: evsmith@uic.edu

Is A Rose A Rose?



Objective analysis of olfactory identification ability in schizophrenia

Kelly Minor

PSYCHOLOGY

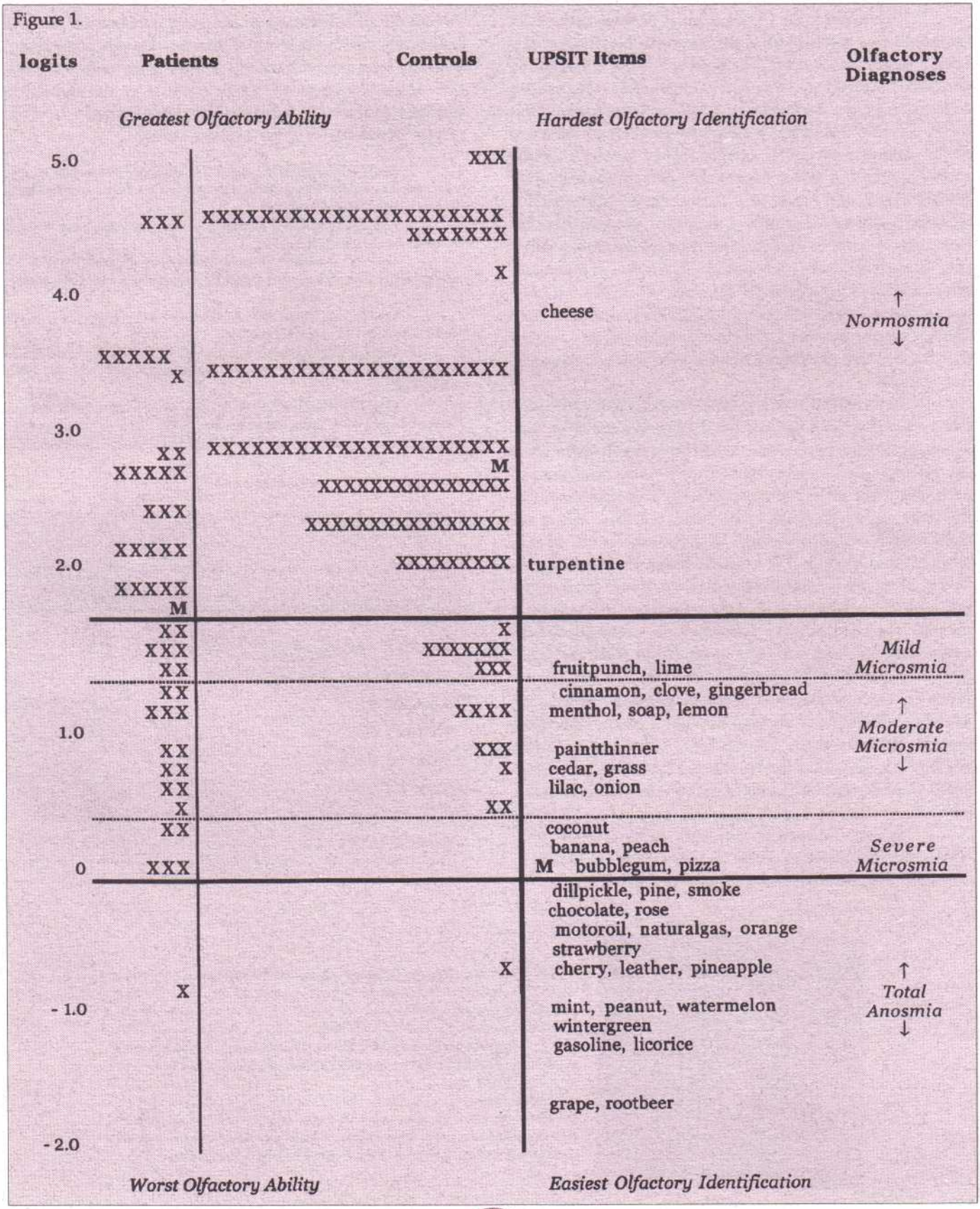
Do schizophrenics smell differently than other people? Indeed, there is evidence to suggest that schizophrenia patients have a unique sense of smell. Despite the fact that schizophrenics have intact olfactory acuity, up to 50% of male patients are reported to be impaired on the University of Pennsylvania Smell Identification Test (UPSIT) (1, 2, 3). This test of olfactory identification ability includes 40 items, each of which presents a scratch-and-sniff patch along with a list of four answer choices. (For example, one item reads, "This odor smells most like: a) chocolate; b) banana; c) onion; d) fruit punch.") Compared to healthy respondents, schizophrenia patients are repeatedly found to demonstrate impaired performance on the UPSIT. These findings, however, are based upon between-group comparisons of raw UPSIT scores and raw scores do not satisfy the basic specifications of measurement. Therefore, we decided to analyze the raw score data matrix from a sample of 54 schizophrenics and 133 healthy participants with the Rasch Model for dichotomous observations (4). The primary goal of our study was to verify that UPSIT items contribute to a single factor with sufficient spread along a discernible line of increasing difficulty to define a recognizable hierarchy of olfactory challenge. We also examined the clinical utility of the UPSIT (i.e., whether UPSIT items separate persons into five distinct levels of olfactory diagnosis as described in the test manual).

Our findings suggest that the UPSIT has succeeded in defining a distinct olfactory identification construct for both schizophrenics and healthy participants. (Item separation indices for the SZ and control groups are 1.70 and 2.49, with corresponding reliability estimates of .74 and .86.)

In **Figure 1**, person ability and item difficulty are expressed in logits and plotted relative to one another (with higher logits representing greater item difficulty and greater person ability). Notice the extent to which the healthy controls (mean, 2.92 logits) manifest better olfactory identification ability on average than the schizophrenics (mean, 2.03 logits). However, the two distributions contain a lot of overlap so that no single cut-off point is available to exclude all controls and also detect most schizophrenics. Even the schizophrenic mean (at about 2 logits) subsumes 26 supposed healthy controls.

Doty (5) reports that the UPSIT "has proved valuable in screening sensory panels in the food and beverage industries, including the water works industry, where a distinction between persons with average or mediocre smell function and those with a more highly developed sense of smell is required." Considering the marked ceiling effect illustrated in **Figure 1**, the utility of the UPSIT in making such a distinction seems unlikely. Our findings show that the average ability of each participant is more than one standard deviation above the average item difficulty. Indeed, nearly half of the controls have an ability estimate above that of the hardest item. Upon examining the item distribution in **Figure 1**, it is clear that the UPSIT does not provide sufficient coverage of olfactory identification ability at the high end. In a region where a majority of both schizophrenic and healthy respondents fall, there are two wide gaps, suggesting that the test does not incorporate enough difficult items to discriminate among higher levels of olfactory ability. Therefore, the capacity of the UPSIT to distinguish average from superior senses of smell is limited.

Figure 1.


logits
Patients
Controls
UPSIT Items
Olfactory Diagnoses
Greatest Olfactory Ability
Hardest Olfactory Identification
5.0
4.0
3.0
2.0
1.0
0
- 1.0
- 2.0
cheese
turpentine
fruitpunch, lime
cinnamon, clove, gingerbread
menthol, soap, lemon
paintthinner
cedar, grass
lilac, onion
coconut
banana, peach
bubblegum, pizza
dillpickle, pine, smoke
chocolate, rose
motoroil, naturalgas, orange
strawberry
cherry, leather, pineapple
mint, peanut, watermelon
wintergreen
gasoline, licorice
grape, rootbeer
Normosmia
Mild Microsmia
Moderate Microsmia
Severe Microsmia
Total Anosmia
Worst Olfactory Ability
Easiest Olfactory Identification

According to the UPSIT manual, persons can be separated into five levels of olfactory diagnosis based upon raw UPSIT score, age, and sex. In **Figure 1**, these standardized cut-offs correspond to the horizontal lines and corresponding olfactory diagnoses are italicized. Rasch person separation statistics measure the UPSIT's ability to discriminate olfactory ability among a particular sample. They are 1.81 for the schizophrenics and .96 for the controls with corresponding reliability indices of .77 and .48. Because the UPSIT score distribution is skewed for healthy participants, standard error has been used in order to illustrate the levels of significant difference in smell ability for each sample. A distance of 3 standard errors implies a significant difference at the 95% confidence level and is indicated in **Figure 1** by solid horizontal lines. Our findings suggest that the UPSIT discriminates three — rather than five — levels of olfactory identification ability among these respondents.

Item difficulty was reported in one study of Parkinson's Disease patients, with patients misidentifying some items (i.e., lemon, pizza, wintergreen, rose, clove) more readily than others (6). Interestingly, our schizophrenic sample did not perceive these particular items to be the most difficult (corresponding item measures are +1.16, -0.04, -1.10, -0.46, +1.16). There are many reasons why patient groups might be expected to differ in terms of item difficulty. For example, target smells differ in intensity, pleasantness, and familiarity of the scent. Further, the test is multiple-choice format with items repeated throughout the test. Therefore, olfactory acuity, attention, memory, and executive function (e.g., perseverative tendency) might each contribute to unique UPSIT profiles for individual patient groups. When item statistics for the schizophrenia group were based upon the item calibrations of the healthy control group, some items showed significant misfit. These items (turpentine, menthol, peach, rose, grape) should be reevaluated for accuracy of presentation and relevance of "wrong" options (e.g., distracter analysis).

In sum, the items of the UPSIT define a single factor of olfactory identification ability and are sufficiently spread to articulate three distinct levels of olfactory identification. Although the UPSIT was not found to separate persons into five statistically significant levels of olfactory identification ability, it clearly separates persons into at least three statistically distinct levels. However, the test is too easy for the majority of respondents and is limited in its ability to discriminate between persons of average to above-average ability.

1. Martzke JS, Kopala LC, Good KP: Olfactory dysfunction in neuropsychiatric disorders: Review and methodological considerations. Biological Psychiatry 1997; 42(8):721-732
2. Doty RL: The Smell Identification Test Administration Manual. Philadelphia, Sensonics, Inc., 1983
3. Doty RL, Shaman P, Dann M: Development of the UPSIT: A microencapsulated test of olfactory function. Physiological Behavior 1984; 32:489-502
4. Linacre JM, Wright BD: A Users Guide to WINSTEPS, Rasch Model Computer Program. Chicago, MESA Press, 1998
5. Hawkes CH, Shephard BC, Daniel SE: Olfactory dysfunction in Parkinson's disease. Journal of Neurology, Neurosurgery, and Psychiatry 1997; 62:436-446
6. Doty RL: The Smell Identification Test Administration Manual. Haddon Heights, New Jersey, Sensonics, Inc, 1995

Correspondence: Kelly Minor, M.A., Northwestern University, 2029 Sheridan Road, Evanston, IL 60208-2710,
(voice: 847-491-5190; k-minor@nwu.edu).

Table 1.

	Controls (N = 54)	Schizophrenics (N = 133)
Mean	2.92	2.03
Standard Deviation (SD)	1.05	1.26
Standard Error of Mean	.09	.17
Real RMSE	.76	.61
Adjusted SD	.72	1.10
Person Separation	.96	1.81
Person Reliability	.48	.77
Item Separation	2.49	1.70
Item Reliability	.86	.74

Kelly Minor is a third-year graduate student at Northwestern University studying clinical psychology (specializing in neuropsychology). Her research primarily focuses on cognitive and behavioral deficits — particularly those believed to involve prefrontal brain dysfunction of schizophrenia patients. The CIC Traveling Scholar Program afforded her the opportunity to take coourses at the University of Chicago — and more importantly, according to Kelly — to meet Ben Wright.

Kelly's hobbies vary according to the time of year. In the summer months, she spends weekend mornings searching garage sales for antiques, and her afternoons refinishing them. In the winter, a perfect weekend includes a day of cooking and baking, then an evening at home watching foreign films.

Adding It Up:

IMPROVED OUTCOMES AND ECONOMIC DEVELOPMENT

William P. Fisher, Jr., Ph.D.
Associate Professor
Department of Biometry & Genetics,
Louisiana State University Medical College, New Orleans

How can Rasch's probabilistic conjoint measurement models contribute to economic development? By deliberately extending into health care and education the crucial role that measurement has historically played in commerce. As anyone who has ever shopped for groceries knows, without fair units of measurement, there is no basis for economic activity. Unfortunately, health care and education lack fair units of measurement for many of their most important outcomes. The lack of stable outcome measures — common currencies for the exchange of quantitative value — in health care and education explains a lot about why it is so hard to know exactly what is obtained per dollar spent in either area, as well as why what is obtained per dollar varies so much across providers.

Functional assessments, test scores, consumer satisfaction surveys, and health status surveys all commonly produce units of measurement that literally do not add up. New technologies exist for correcting that situation and thus for enhancing the economic impact of health care and education.

In a nutshell, Rasch measurement is playing a crucial role in the founding of a new kind of organization, a developmentally-attuned organization that learns from the cumulative experience of its members. These organizations deliberately and scientifically measure learning and experiment with it, becoming organizations that coordinate their own evolution with the evolution of their members and partner organizations. This literal "co-ordination" will be the impact of units of measurement that do not vary in their size or order (by more than a known amount of error) depending on which brand instrument they are read off, where or when they are used, or by whom. With these instruments in systematic use throughout these organizations, even spontaneously occurring consistent variation in clinical or educational outcome measures will support better understanding and improved treatment and teaching effectiveness, quite apart from controlled experimentation.

The key is to put useful and meaningful information into the hands of those responsible for outcomes, the providers and consumers, teachers and students, in a form designed to be as developmentally well-targeted, and so intellectually accessible, as possible. The history of science shows over and over that widely recognized, interpretable, consistent, and reproducible effects and phenomena are essential to new understanding. The historian of science who coined the phrase "big science" once remarked that "thermodynamics owes much more to the steam engine than ever the steam engine owed to thermodynamics," and that "the chemical revolution resulted much more from the technique of the electric battery than from the careful measurements or new theories of Lavoisier" (Price 1986, pp. 240, 248). Computer adaptive testing and health assessment will be to medical education and health care what the steam engine was to thermodynamics and the electric battery was to chemistry because these technologies will provide universal access to consistently reproducible and interpretable quantitative information.

Those who are among the first to understand this will have the lead in making the new techniques pay off. The payoff is going to come in the form of evidence that supports or contradicts the effectiveness or efficiency of treatments, teaching, policies, or skill levels. Simply by placing fair and universally-recognizable units of measurement in the hands of people on the front lines of health care and education, many of those people will immediately take responsibility for the outcomes of their practices in ways that they never could before. When they are readily able to see the effect of variations in their treatments on outcomes, they will come to understand what they did not understand before. When they can compare the results of their interventions with the results of the clinic down the street, the practitioner across the hall, and the group across the country, they will either take new pride in their work or want to know how to do better.

And the comparisons, the reflection, and the decision to take action will not be a cumbersome, time-consuming, expensive process of data gathering and analysis. Instead the data system will already be in place. The relevant measures will all be expressed in a common quality assessed and monitored unit. In the same way that steam engines and electric batteries

permitted observations of thermodynamic and chemical effects to be routinized and generalized, so too will calibrated measures of educational and medical outcomes permit the routine and general observation of teaching and treatment effects. Only when such observations can be expressed in common units of measurement will they be able to play a significant role in the larger conversations of professionals actively engaged in learning and sharing their learning. Richer community lives for those impacted by measurement can be effected only when ways of sharing richer common unities of meaning are provided. Rasch measurement practitioners are discovering, inventing, and creating those unities.

Some Historical Background

As Ben Wright likes to point out, many historical documents, including the Bible and the Magna Carta, specify particular units of measurement as a standard in order to promote a common currency for the exchange of value. Looking back over the course of time we see that empires and political alliances bring about large economic communities that share much in the way of measurement standards, and that political fragmentation is associated with wide variation in measuring units. Before the French revolution, every town in Western Europe had its own system of weights and measures; Napoleon adopted the revolutionaries' metric system as a means of unifying the empire, with the effect of stimulating trade across a wide region.

Today, health care and education are like Western Europe before the French Revolution. There are wide regional variations in treatments, outcomes, and costs. Outcome measurement has been identified as a potentially useful means of overcoming some of the unwanted variation and of making treatment effects comparable. Existing quality of life and health status measures fall far short, however, of being equivalent to Napoleon's unified metric system.

Flaws in Current Practice and New Ideas for the Future

To date, virtually all efforts aimed at measuring outcomes have merely added more levels to the Tower of Babel as the proliferation of new instruments has brought along with it a proliferation of new units of measurement. Each different instrument has its own particular questions and its own rating response format, meaning that the sum of the ratings means something different for each instrument. One survey has 10 items and 3 response categories, for a 10-30 score range, and another one, intended to measure exactly the same thing, has 20 items and 6 categories, for a 20-120 score range. Scores from the two instruments plainly do not correspond. And even if complex statistics were used to establish correspondence, the two instruments still would not be shown to measure the same thing; the scores would still be nonlinear, nonadditive, and ordinal instead of linear, additive, and interval; the measures would not be accompanied with error and data quality estimates; and every respondent would have to answer every question on the instrument of choice for even the appearance of comparability.

What if, however, the two instruments had been calibrated to measure in one quantitative metric? What if the two instruments were used in two different hospitals in two different clinics seeing the same kind of clients? Using the new measurement technologies, given sufficient data quality, outcomes could be compared across the two clinics even if clients routinely skipped questions or if the instruments were in fact adaptively administered, so that clients were asked only those questions relevant to their condition.

Similar scenarios involving the comparison of test results across universities, course sections, or from year to year could be imagined in the educational arena. When universal access to universally interpretable and comparable educational measurement information is available, a new economy of educational effectiveness studies will be created.

Our goals ought to be 1) that each variable measured by means of client self-report satisfaction and quality of life surveys, clinician-administered functional or performance assessments, or psychological and educational tests is calibrated to a reference standard, 2) that the quantitative units of the majority of instruments in use for measuring each of those variables be traceable to that standard, and 3) that metrological systems be put in place for monitoring the quality of the instruments and the measures. Only when these goals are achieved will there be a basis for trading in and banking on a common currency of health and educational value. Only when these generalized measuring units are brought into the clinics and the classrooms and put into the hands of the care providers and consumers, and the teachers and the students, will people on the front lines of health care and education have the information they need to take responsibility for the outcomes of their efforts.

Suggested Plan

In the first five years of this plan, organizations interested in advancing a broad-based measurement agenda should publicly establish themselves as being aware that new management efficiencies could be provoked by the creation of unified systems of measurement for educational, psychological, and health care outcome variables. These organizations should make all of their faculty, students, staff, customers, etc. aware of several points, directing the early adopters to publicly available bibliographic resources:

1) that units of physical measurement (meters, grams, volts, ohms, degrees centigrade, etc.) do not exist in nature but are the result of ongoing
   a) experimental research establishing a convergence of results across samples, labs, instruments, and other variables according to strict mathematical data re-

quirements, and

b) efforts deliberately aimed at creating and maintaining measuring units as the common currency for the exchange of quantitative value;

2) that the only reason why there are no unified metrics for psychosocial variables is that no one has set out to create them with the right tools (perhaps because of the heavy computational burdens and lack of accountability demands);
3) that long-established measurement theory and data analysis techniques are available to help create unified metrics;
4) that applications of this theory and these techniques are establishing the expected experimental convergence of results across samples, labs, instruments, etc.;
5) that increased economic pressure and accountability in health care and education demand easily understood and comparable outcome measures, measures that can come into being only when sufficient attention is paid to instrument design and user training;
6) that it is no longer necessary to force people to adapt to the needs of measurement technologies, as it is now possible to adapt the measurement technology to the needs of people;
7) that examinees and survey respondents need not answer one single set of questions to be meaningfully measured in a common quantitative unit;
8) that measures can therefore be made comparable across classes, years, clinics, cultures, etc., even when tests and surveys are not identical, opening up vast new possibilities for understanding variation in learning and health;
9) that organizational growth of the kind we envision will most readily occur in a learning environment that recognizes that intellectual development does not stop at any age, but that adults can progress through as many marked transformational stages of development between 21 and 70 as they did between birth and 21; and
10) that part of the general mission of educational and many other kinds of organizations must be to provide an environment that supports continuing intellectual development to those they serve, including their own employees, and to take differences in reasoning attributable to developmental variations into account in educational, clinical, and managerial decision-making.

In the second five years, the paradigm-shifting organizations should establish themselves as the world leaders in the calibration and use of unified metrics in human resource management, health care, and medical education. They should set an example for the world to follow in its use of unified measures, linking up with other universities, hospitals, employers, schools, and government agencies locally, regionally, and globally to form the networks through which unified metrics for psychosocial variables will be created and maintained. Someday there will be a meaningful, useful, and quality-assessed quantitative metric for each psychosocial variable (clinical competence; physical function; consumer satisfaction; knowledge of anatomy, spelling or mathematics; environmentally sound behaviors; etc.) we're interested in, just as we have for each physical variable (meters, grams, volts, etc.). Organizations employing Rasch measurement practitioners could, should, and are playing a big role in making that happen.

The second five years should focus 1) on disseminating calibrated measuring instruments traceable to universally recognized reference standards; 2) educating faculty, students, clinicians, patients, employees, employers, etc., on instrument use, with demonstrations of each of the above 10 points; 3) on technical aspects of instrument design and calibration; and 4) on the needed information systems.

Overall effectiveness of the measurement program could be evaluated by surveying participants as to how much they are learning about their practices now, and surveying them again periodically as instruments come on line.

Summary and Conclusions

The economic need for common units of measurement greatly predates science. Economic development is impossible unless we can estimate amounts of value in a way that does not depend on the particulars of the measurement process, such as who is using which brand instrument where and when. Recent advances in measurement theory and in computerized information technologies support the emergence of new kinds of learning organizations capable of deliberately evolving in the direction of enhanced efficiency and effectiveness. Rasch measurement practitioners have a unique opportunity to help shape the organizations that will create new health care and educational economies. Carpe diem!

Reference

Price, D. J. d. S. (1986). Little science, big science...and beyond. New York: Columbia University Press.

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.

Continuum of Care

Measuring Medical Rehabilitation Outcomes

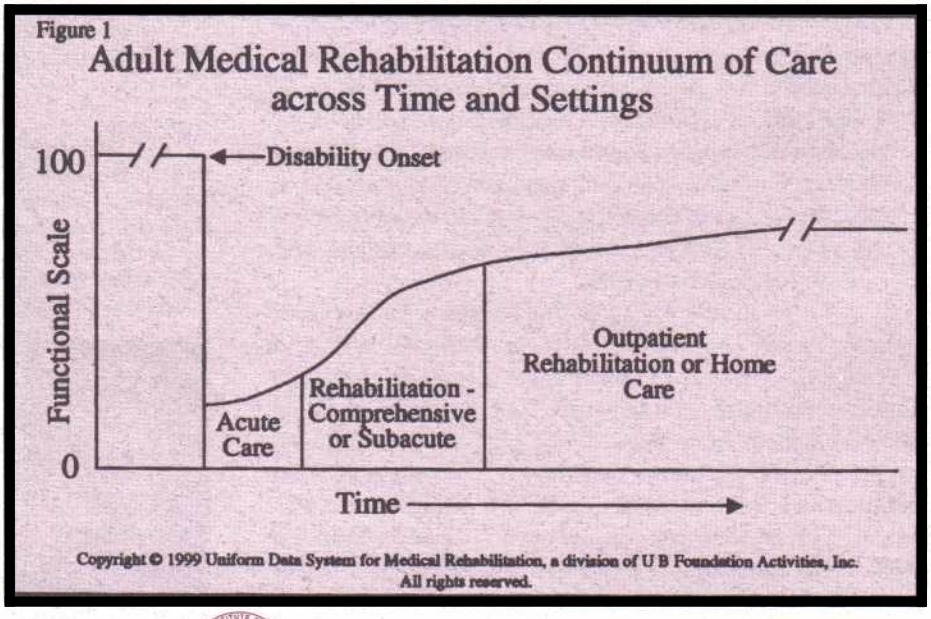
Carl V. Granger, M.D.

Health care is delivered in an array of settings. Patients start in acute care hospitals then may further recuperate through a continuum of care. This is especially true for patients requiring medical rehabilitation services because of an acquired disability such as stroke, spinal cord injury, or orthopedic surgery. The increased quality of medical technology stabilizes more patients after medical crises. But they may be left with deficits in activities of daily living, such as walking, eating, dressing, or communicating. Restorative patients have the potential to improve and, therefore, may benefit from post-acute therapies to regain lost function. The goal is for these patients to attain a level of independence that enables them to return to their former lives. Under the direction of a rehabilitation medicine physician, they may receive physical, occupational, vocational, speech, and recreation therapies. The therapeutic settings that comprise the continuum of care are changing dramatically. Restorative patients may move from acute care hospitalization to comprehensive medical rehabilitation units within hospitals or to freestanding rehabilitation hospitals. Or they may move to subacute care in a rehabilitation facility, hospital, or skilled nursing facility. Or they may receive outpatient care, or home care, or adult day care. Restorative patients take different paths to recovery. All services, regardless of setting, should be interrelated and coordinated. These services and settings should be seamless, with no overlaps or gaps. This is the ideal of a continuum of care. (See Figure 1.)

To realize this ideal, we need the ability to measure and manage outcomes and predict which types of patients will benefit most in which settings, at which times during their illness, the duration of services, and costs. Managed care capitation is forcing health care institutions to learn how to achieve cost-effective outcomes. These outcomes are determined by measuring patient improvement, length of stay, patient satisfaction, and costs.

Measuring medical rehabilitation outcomes has been the mission since 1987 of the Uniform Data System for Medical Rehabilitation (UDSMR), located in the School of Medicine and Biomedical Sciences at the State University of New York at Buffalo. UDSMR has developed a family of outcomes measurement tools that the rehabilitation industry uses on a daily basis. Measuring outcomes across the continuum of care — time and settings — is a major UDSMR goal. UDSMR intends to maintain beginning-to-end care information on patients, which will enhance predictability of outcomes along the various rehabilitation paths. All UDSMR instruments are modeled on the Functional Independence Measure (FIM instrument) developed by UDSMR, and used internationally in medical rehabilitation and subacute care settings. The FIM instrument is an 18-item, seven-level scale assessing the functional status of patients with disability. Trained staff administers the instrument. Patients are observed before, during, and after their therapy regimens and rated, performing 18 motor and cognitive activities of daily living. Ratings for each item range from total assistance needed, represented by one, to complete inde-

Figure 1



REHAB MEASUREMENT

Figure 2

FIM™ instrument

FIM items

Motor Items

Self-care
A. Eating
B. Grooming
C. Bathing
D. Dressing - Upper Body
E. Dressing - Lower Body
F. Toileting

Sphincter Control
G. Bladder Management
H. Bowel Management

Transfers
I. Bed/Chair/Wheelchair
J. Toilet
K. Tub/Shower

Locomotion
L. Walk/Wheelchair
M. Stairs

Cognitive Items

Communication
N. Comprehension
O. Expression

Social Cognition
P. Social Interaction
Q. Problem Solving
R. Memory

FIM levels

No Helper
7 Complete Independence (Timely, Safely)
6 Modified Independence (Device)

Helper
Modified Dependence
5 Supervision (Subject = 100%)
4 Minimal Assist (Subject = 75%+)
3 Moderate Assist (Subject = 50%+)
Complete Dependence
2 Maximal Assist (Subject = 25%+)
1 Total Assist (Subject = less than 25%)

pendence, represented by seven. Ratings in between represent levels of assistance needed by the patient, provided by a helper or assistive devices. (See **Figure 2.**) The goal is to bring patients to independence, so that the burden of care on family and society is minimized. Higher ratings represent less severity of disability and less burden of care. Lower ratings represent more severity and more burden of care. Patients are rated at admission to a rehabilitation program, during rehabilitation, at discharge, and follow-up. Functional improvement is analyzed in the context of time spent in each setting, resources expended, patients' ages, and classification and severity of impairments (called Function Related Groups, or FRGs, developed with the University of Pennsylvania). FIM-FRGs make predictions about time and resource utilization. Subscribers to The FIM SystemSM send data to UDSMR and receive reports on their program and an aggregate comparison report that includes other programs in the region and around the country.

The Alpha FIM instrument, a six-item abbreviated version of the FIM instrument, measures a patient's functional status during the first 72 hours of acute care hospitalization. Its greatest value is in triage: determining the patient's next appropriate care setting and pinpointing the earliest opportunity for entering that setting. The Alpha FIM instrument measures eating, grooming, bowel management, toilet transfer, expression, and memory.

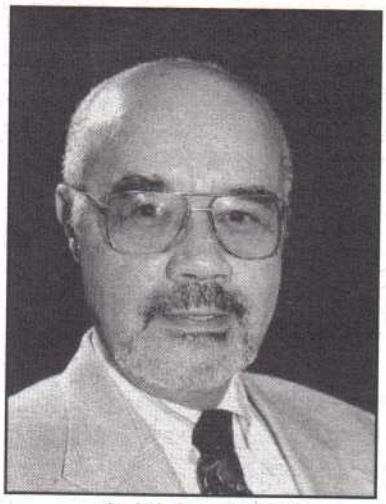
The LIFEware System℠ measures the functional status of individuals who receive outpatient rehabilitation services. Two major impairment groups are addressed: musculoskeletal conditions and neurological conditions. Outpatient status is assessed in the domains of physical functioning, affective well-being, cognitive functioning, and pain experience. Some FIM instrument items appear in the LIFEware survey instrument, but unlike the FIM instrument, which is administered by trained clinicians, LIFEware is usually self-administered in the outpatient setting. LIFEware is Internet-driven. Continuum of Care Reports are available to LIFEware subscribers, linking U.S.-based inpatients and outpatients by their social security numbers, showing number of outpatient visits for any series of treatments. Though designed for adults, LIFEware may also be useful for children seven years and older. However, organizations serving pediatric populations are better served by UDSMR's pediatric instrument, the WeeFIM™ instrument. The WeeFIM System℠ assesses children and young adults, 6 months to 21 years old, in three domains: self-care, mobility, and cognition, and is adjusted for age. It is useful for children with acquired disability, congenital disability, and developmental delay. Like the FIM System, the WeeFIM System has a national database. For children the setting continuum is different because most pediatric disabilities are congenital, so children with disabilities seldom begin rehabilitation in an acute care hospital.

The HomeFIM instrument measures the functional status of patients receiving rehabilitation intervention in their homes. This system contains items from the FIM and LIFEware instruments that measure function levels, units of service, and patient satisfaction.

The FIM System, HomeFIM, and the WeeFIM System are used for accreditation purposes with the Joint Commission on Accreditation of Healthcare Organizations.

In addition to serving its subscribers with data analysis and reporting, UDSMR maintains a national data repository for research purposes of three million case records from 1,400 facilities around the world, 1,200 of which are in the United States. There is no recognized prototype for a continuum of services system in medical rehabilitation. UDSMR research is synthesizing clinical and administrative knowledge. For the restorative patient, this could result in better rehabilitation care,

a quicker return to the community, and improved quality of daily living. An important feature in feedback reporting reflecting a patient is status at each assessment encounter and cumulative scores on a quarterly basis. An ongoing database enables periodic exploration for factors that contribute to best practices and predict the likelihood of future events. UDSMR offers medical rehabilitation measurement instruments from youth to old age, across the continuum of care. We are concerned with methods that are feasible in multiple clinical settings, which provide relevant feedback to providers, and which are compatible with principles of measurement that support meaningful interpretation of data.



Carl V. Granger, M.D

Carl V. Granger, M.D. is Professor and Chairman of the Department of Rehabilitation Medicine and Director of the Center for Functional Assessment Research at the State University of New York at Buffalo. Dr. Granger is a graduate of Dartmouth College and New York University School of Medicine. He is certified in Physical Medicine and Rehabilitation and Electrodiagnostic Medicine.

Dr. Granger is past president of the American Academy of Physical Medicine and Rehabilitation and of the International Federation of Physical Medicine and Rehabilitation. He served on the Advisory Committee of the National Center for Medical Rehabilitation Research, NICHD, NIH.

Dr. Granger has over 150 publications. His interests and research are in the development and use of measures of disablement and quality of daily living. This includes physical, mental/emotional, and social functioning in order to evaluate outcomes of medical rehabilitation. He is one of the developers of the Functional Independence Measure (FIM) instrument and the Uniform Data System for Medical Rehabilitation (UDSMR).

Dr. Granger was awarded the Krusen Award from the Academy of Physical Medicine and Rehabilitation and was its 30th Annual Walter J. Zeiter Lecturer. He received the Elizabeth and Sidney Licht Award for Excellence in Scientific Writing and is Fellow (Honoris Causa) in the Australian Faculty of Rehabilitation Medicine of the Royal Australian College of Physicians. Dr. Granger is a member of the executive committee of the RRTC on Functional Assessment and Evaluation of Rehabilitation Outcomes. He is active with Habitat for Humanity. He can be reached at granger@acsu.buffalo.edu.