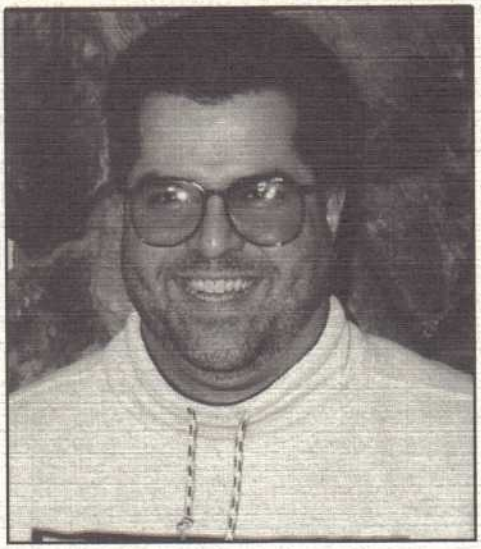
Objective Analysis Of Golf

Patrick Fisher, M.A.



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

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

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

Table 1 - Persons Measurement Report

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

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

Table 2 - Day Measurement Report

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

Reliability 0.92

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

Table 3 - Holes Measurement Report

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

Reliability 0.92

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

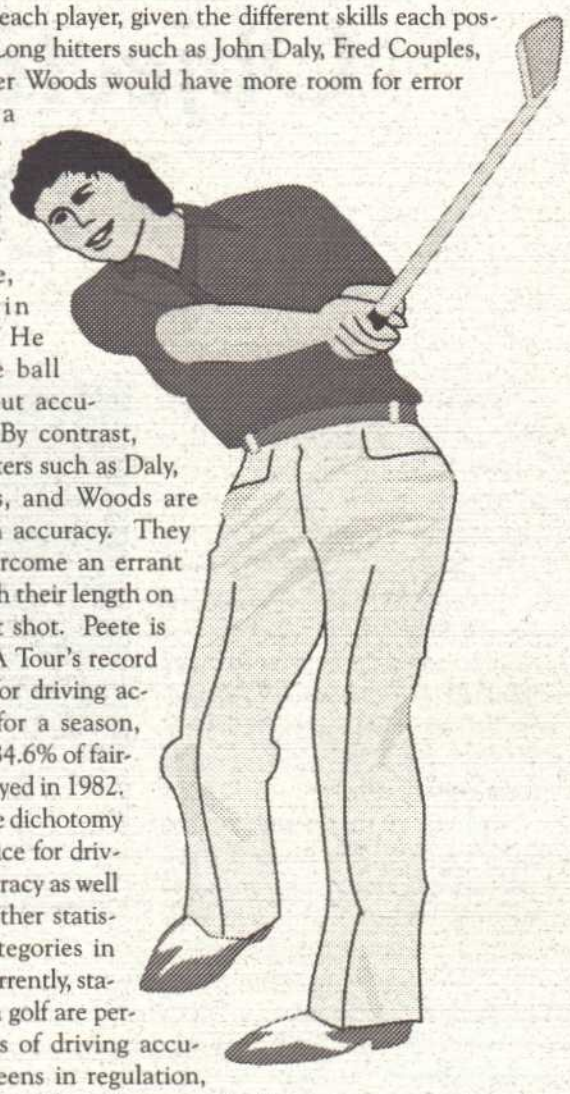
Table 4 - Misfitting ratings

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

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

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

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



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

Patrick B. Fisher, MA

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

Assessment:

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

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

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

What Is It?

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

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

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

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

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

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

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

Why Do We Need It?

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

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

definition is included as to what that means.

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

How Do We Use It?

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

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

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

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

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

Roy Berko

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

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

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

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

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

Machiavelli

Public Speaking Assessment for College Students

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

William W. Neher

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

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

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

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

Deborah Jean Grew
Director, Computerized Public Speaking Assessment
Butler University

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

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

Her favorite travel spots are Maine and Cape Cod.

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

Max Plank

Student Progress? Prove It!

Donna Surges Tatum, Ph.D.

Course Goals

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

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

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

To enhance delivery skills

To teach methods of organization and critical thinking skills

To increase confidence.

Research Questions

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

Data Description

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

Assessment Issues

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

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

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

1. What are appropriate expectations?

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

2. Are the evaluation instruments sound?

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

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

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

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

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

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

4. How can we compare results?

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

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

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

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

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

Results

Units of Measure

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

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

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

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

Raters

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

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

Items

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

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

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

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

At Level 3 the visual impression of the speaker, word

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

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

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

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

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

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

Speech Results

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

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

ITEM MAP

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

Speaker Improvement - 2 Speeches

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

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

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

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

Speaker Improvement - 3 Speeches

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

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

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

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

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

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

Speaker Improvement - 4 Speeches

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

Speaker Improvement - Advanced Class

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

Rater Consistency and Speaker Ability

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

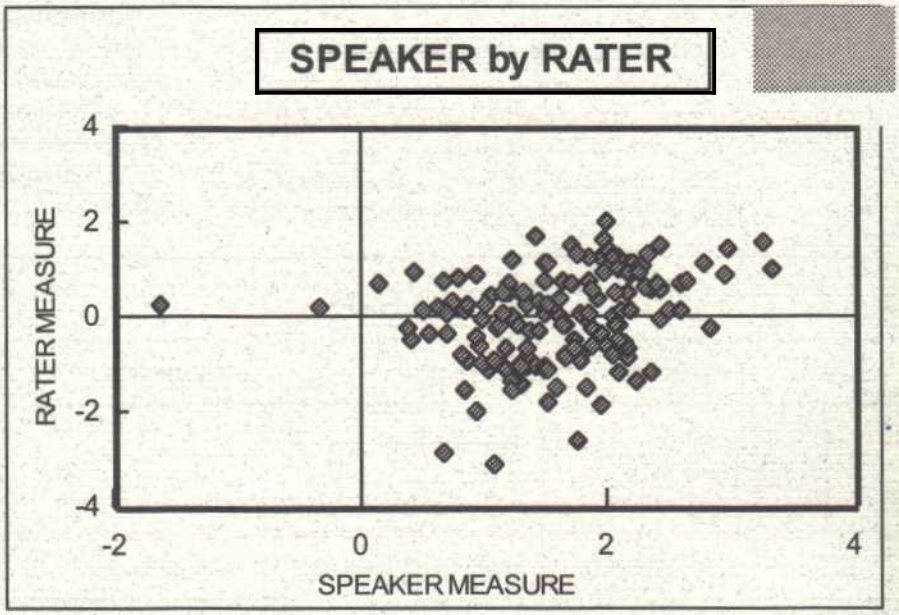
A rule of thumb is to look closely at any response pattern with a mean square of more than 1.4, or a standardized fit over 2. When this occurs, a red flag waves in the researcher's mind, and a close examination of the data is warranted to determine the cause of the misfit. It may be that the rater is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

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

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

Rater Severity and Speaker Ability

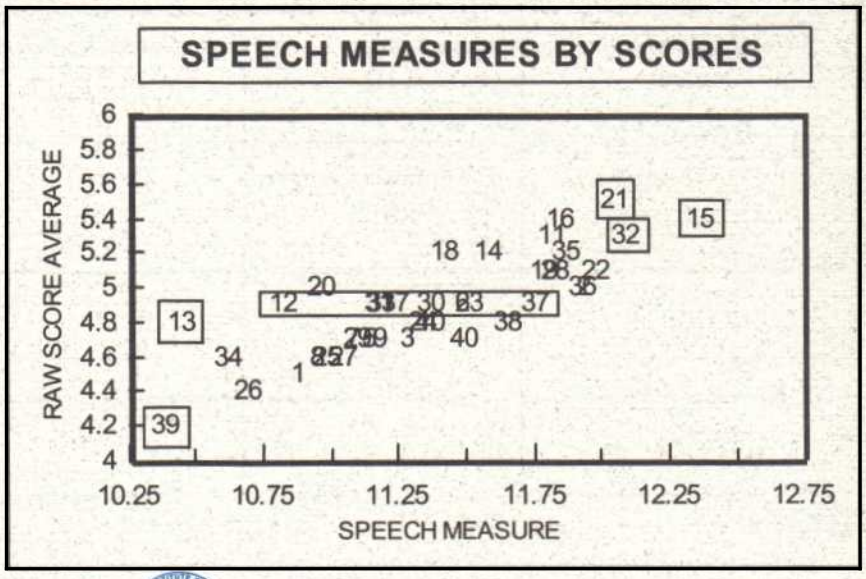
The graph below shows there is not a clear relationship between a person's severity as a rater and their ability as a speaker. Some excellent speakers are easy raters, and some poor speakers are quite severe.



Measures and Raw Scores

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

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



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

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

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

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

Discussion

Meaningful Measurement Results

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

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

This study also demonstrates:

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

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

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

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

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

Let Us Put Your Back Into Action

THERAPY PROVIDERS OF AMERICA

Physical, Occupational and Speech Therapy

- *Oncall P. T. 24 hours*
- *Transportation*
- *W/C Accessible Clinics*
- *State of the Art Equipment*
- *Home Calls*
- *Accept Medicare*

Speech Therapy
Occupational Therapy
Hand Rehabilitation
Pediatric Care
Work Hardening
Back School

Clinics Located in

Oakbrook Medical Center

Oak Lawn, Illinois

Phone 1 800 403-7279
Fax (708) 229-0084

Health Care Outcome Measurement

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

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

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

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

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

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

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

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

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

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

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

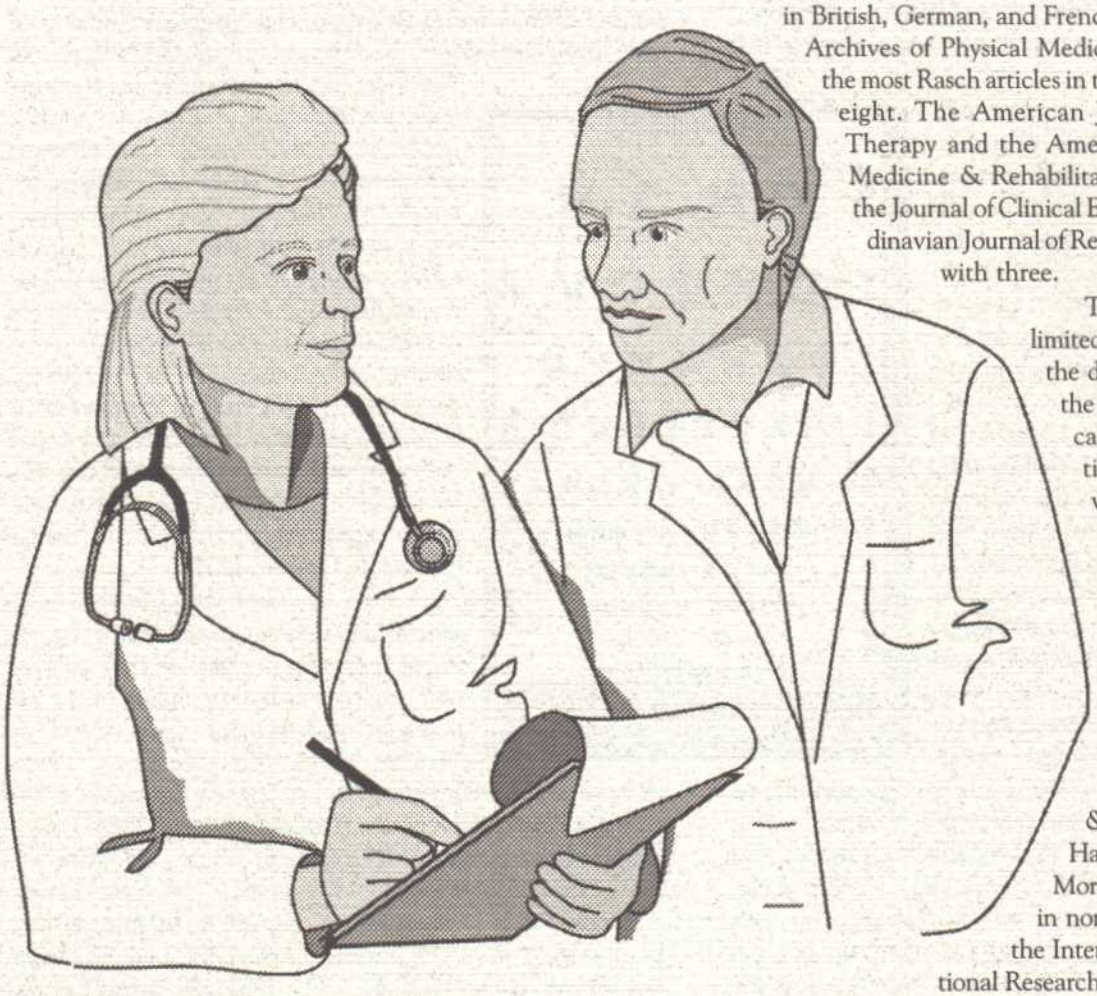
With Harvey as editor and the submission of articles reporting advanced measurement research employing functional assessment instruments, the Archives of Physical Medicine and Rehabilitation became the leader in rating scale measurement and practice among health care publications. A key moment arrived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

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

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

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

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



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

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

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

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

Instantaneous Measurement and Diagnosis

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

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

Better Measurement

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

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

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

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

Calibrating the Measurement System

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

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

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

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

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

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

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

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

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

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

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

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

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

Constructing the KeyFIM

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

Figure 1. KeyFIM data collection and analysis sheet.

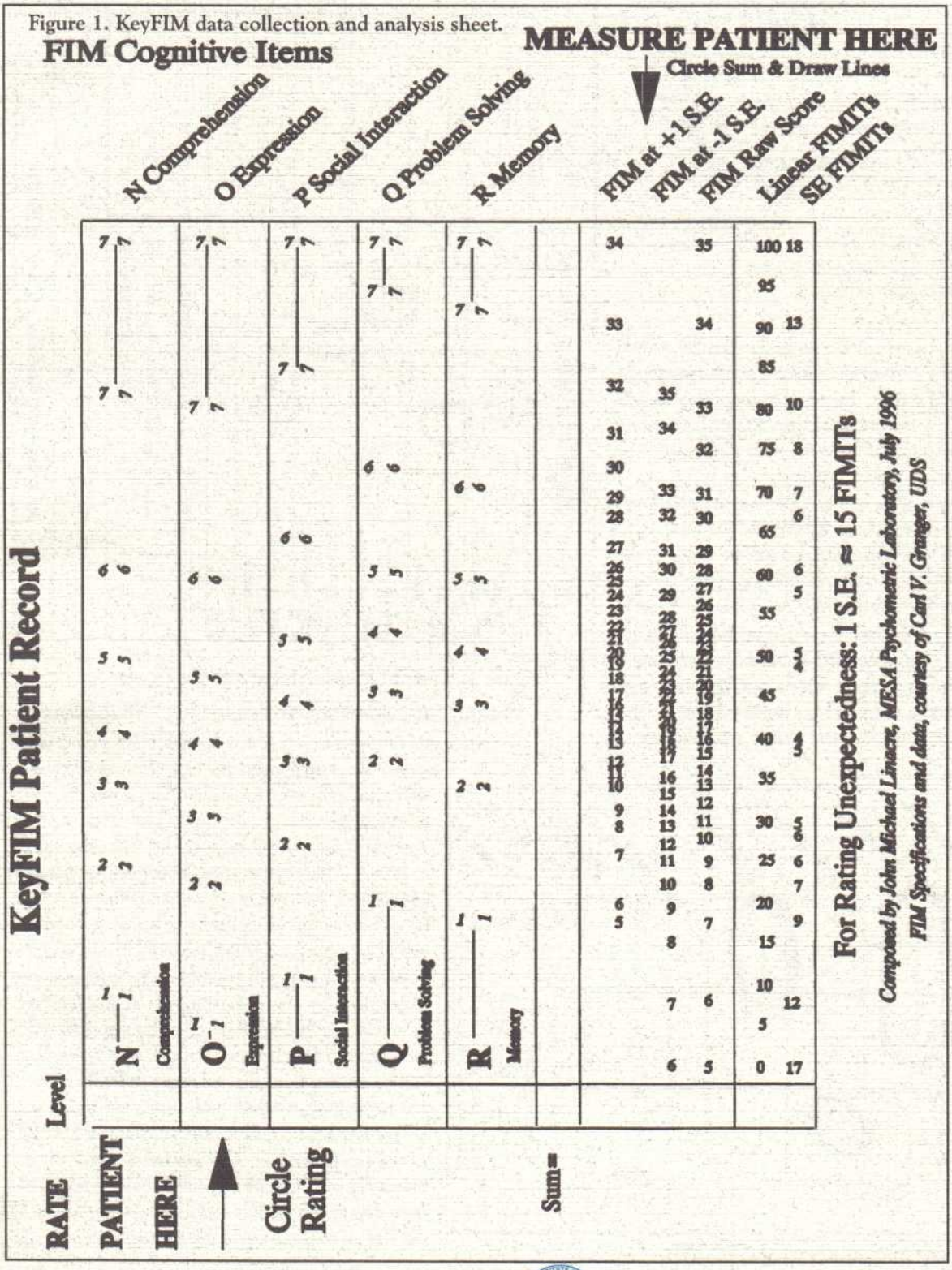


Figure 2 shows an actual patient record from IGC group 13, "Other Impairments." The KeyFIM has been turned on its side and the FIM levels recorded for each of the 5 cognitive items: 3 on item N. Comprehension, 3 on item O Expression, etc. The FIM ratings total 16. The corresponding levels are circled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

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

REHAB MEASUREMENT

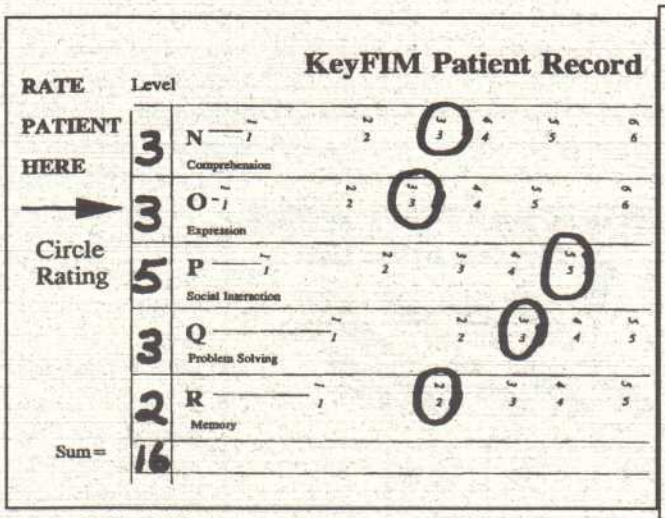


Figure 2. KeyFIM data collection.

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

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

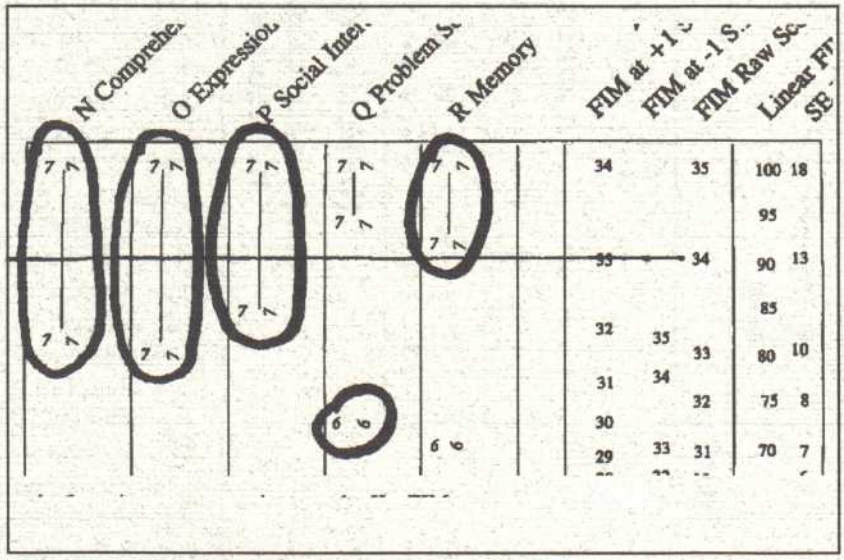


Figure 4. Locating extreme ratings on the KeyFIM.

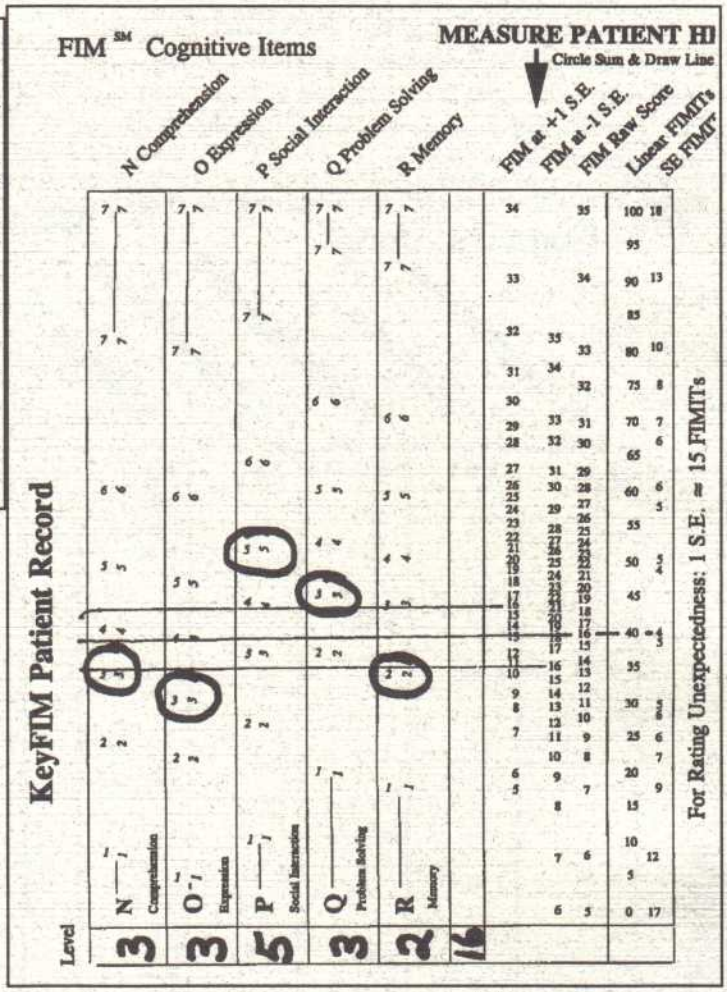


Figure 3. KeyFIM Measurement and Fit Diagnosis.

Instantaneous Measurement and Diagnosis

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

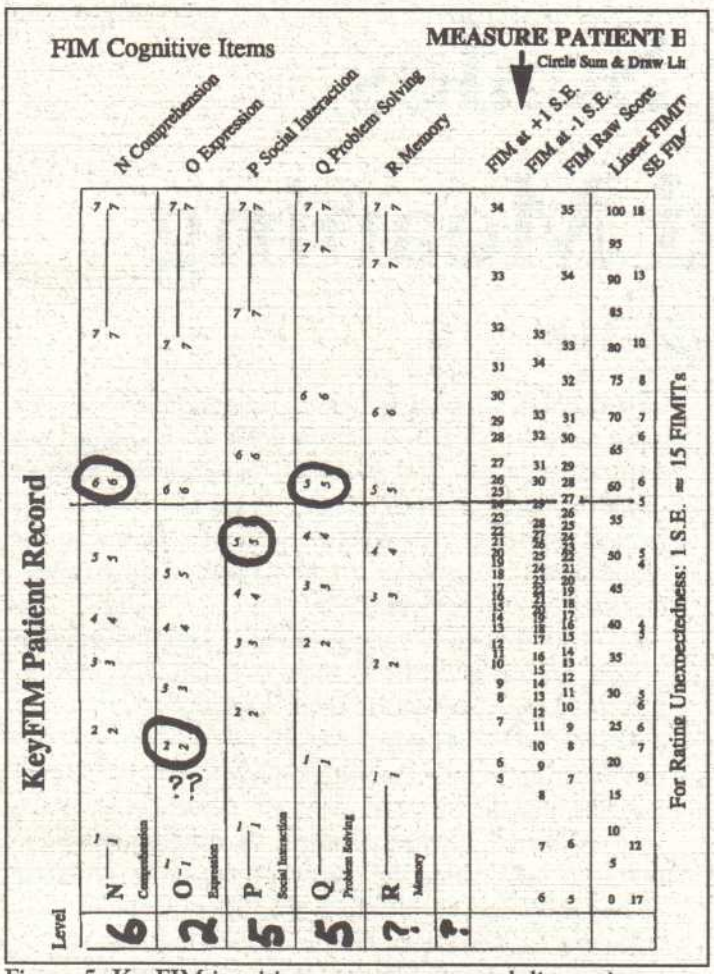


Figure 5. KeyFIM intuitive measurement and diagnosis.

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

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

Conclusion

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

Notes:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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