STANDARDS
Refining Measurement

Eric C. Haughton

Introduction

People working within Precision Teaching are using some of the most accurate, sensitive and valid formative data available in the Human Services area. There are several reasons supporting this fact, including the pioneering work of B. F. Skinner in the use of frequency as basic data and the work of Og Lindsley in monitoring successive frequencies of thousands of human acts. Many other contributing people could also be mentioned. However, being aware of your Precision Teaching mentors and their geneology will create a reasonable list.

Some of us in teaching or research run into confrontations with our historical measurement precedents. Test styles have been so rigid for the past 40 years, that new approaches, however clear and operational, are often placed on the defensive. My own experiences as one of the early field advisors and supervisors causes me to be quite sensitive and practiced in discussing this interesting, while complex area. In the past few years I've been developing an overhead transparency and handout to attempt to explore this labrynthian, nether region. Furthermore, we continually wish to refine our information base, so discussion along with comparisons may unearth other significant factors.

Contrasts

Repeated measurement or monitoring forms a cornerstone of a different foundation of information than that of traditional or commercial testing. Two reasons we monitor performance are to chart changes and to forecast change on the personal level. We relate the individual's charts to group or other reference data. Testing attempts to relate group data to static individual data. Actuarial data (used by insurance and testing companies) cannot forecast individual's outcomes.

Precision Teaching practitioners monitor individual and programme-related concerns. Commercial tests, cover the waterfront, including a wide spectrum of topics, in order to meet market and administrative needs, not those of individuals.

Both measuring systems attend to the two major **Quantities:** Quantity 1 is temporal (calendar and interval) and Quantity 2, the content of the performance. Testing obscures the frequency data inherent in all standardized tests while frequency is a consistent unit of our information.

Both systems work to ensure the accuracy of their data. Precision Teachers break performance into significant packages to explore and to meet individual needs and characteristics accurately and precisely, such as corrects, skips and learning opportunities. Testing generallly relates only to accuracy and acceptability.

Measuring—from Test to Monitoring

Let us now go through Figure 1 with my brief comments. Each of you will have personal experiences to relate to, so mine are designed as stimulants, telegraphic. We'll hit the high spots and clarify some of the hot spots.

Commercial/Personal: Standardized tests often commit to multiple-choice and machine scoring formats for largely economic considerations. These formats can be intimidating and distracting to both behavers and managers. We strive for a fully informed team involving usual behaviours, high comfort and trust levels. Data on my personal pinpoints are for, and belong to me personally.

Minifeedback/Maxifeedback: In the worst testing situations even concerned teachers do not learn results. Behavers who chart regularly receive maximum, immediate, feedback while managing their own projects. Behavers are operating as self managers and resource seekers.

Average/Proficient: Standardized tests relate your performance to the mean of your peers. Suppose you are in first grade, and the mean peer-norm is 50 words correct per minute on oral reading. Is 50/minute competent, fluent or proficient?

We use different frames of reference depending on the behaver's interests, desires and needs. We may ask for a personal aim at the start: "better than I was." We may use some peer data, and in the final analysis we owe it to each behaver to determine levels that will ensure Retention, Endurance and Application of their learnings. Many of us recognize this topic as deserving immediate study. Since we are a "Nation at Risk" we need to determine and implement education based on substantial proficiency levels. Few decision guidelines exist. What performance levels do you use when deciding on new phases? How much is enough?

Fail/Support: If you can't answer an item on a test, you fail the item. One of the classic IQ items is "What is Mars?" Robin answers, "Candy bar." She fails the item. We structure monitoring to support Robin through changes. There are supports for productive change in

Early Childhood Education
Loyalist College
83 11 10 (M-MT5)

Figure 1

MEASURING

TESTING (STATIC)

- PRE-PACKAGED FORMAT
- USUALLY NO STUDENT FEEDBACK
- RELATES ONLY TO **PEER-NORMED** GROUP
- CAN FAIL - "OBSCURE" CRITERIA
- SNAPSHOT
- CAN'T FORECAST - ONLY RELATE GROUP STATIC NORMS TO YOU
- UNRELATED TO CLASSROOM OUTCOMES AND GOALS - TENDS TO BE INSENSITIVE AND HETEROGENEOUS
- MULTIPLE CHOICE
- QUALITY ONLY (PACE IMPLIED) WITH GRADE/AGE LEVEL CONTENT
- DIFFICULT TO VERIFY (VALIDATE) OURSELVES
- GRADE/AGE LEVEL TRANSLATION TO **METAPHORS**

MONITORING CHANGE (DYNAMIC)

- CLASSROOM TOPICS AND CONCERNS
- STUDENT RECEIVES FEEDBACK
- RELATED TO COMPETENCY - RETENTION ENDURANCE - APPLICATION - PERFORMANCE STANDARDS (REA/PS)
- FAILURE FREE - INFORMATION
- CONTINUOUS PROCESS
- FORECASTING POSSIBLE: REPEATED SERIES NECESSARY - CHARTED TO BE EFFECTIVE
- DIRECTLY RELATED TO EDUCATIONAL GOALS - VALID, RELIABLE, SENSITIVE, HOMOGENEOUS
- CONSTRUCTED RESPONSE
- ASSESS PACE AND QUALITY, AS WELL AS RATE OF CHANGE
- CAN BE VERIFIED (EMPIRICAL) IN EACH CLASSROOM
- DIRECT REFERENCE TO COMPETENCE (REA/PS) STANDARDS

Precision Teaching procedures and techniques (Feel Better, Robin?)

Snapshot/Continuous: Testing, even pre-post, are one shot events. Whereas we check performance repeatedly based on regular calendar cycles.

Now/Forecasts: We cannot forecast from a single observation---not navigators, not weather folks, not medics, not ETS, not behavers, not managers, not you and, certainly not me! Therefore snapshot, commercial tests offer static hints about a person's strengths and needs. Since we don't know the rate of change, we can't estimate the necessity of intervention or the intensity of intervention required. T en day screening data improves people at x1.3 M/m/week, on the average--some more, some people less. We have learned not to project a flat line from initial frequencies, an unfortunate, incorrect assumption in current testing and statistical approaches. Slope is one of our power-pieces to understanding measurement and individuals. This should make you feel good as well as proud of our steadfast group.

Unrelated/Relevant: U.S. law 94-142 requires that measurement relate to behaver's programmes and goals. Goodbye IQ. Adios, traditional diagnostic and labeling testing. Au revoir to heterogeneous test sections. Hello to valid, reliable, usually homogeneous items with **SENSITIVITY**! The fact that our data are sensitive is worth more consideration, so see the next thrilling installment of this column.

Prompt/Produce: Prompted, test-taking behaviour (multiple choice format) differs from normal, performance ecology. Each of you know several anecdotes about people who have guessed their way to "success" in prompted tests. Success? Monitored performance is similar to real-life production, often requiring multiple, compound learning channel sets. This is in marked contrast to commercial testing's slavish use of See/Select-Mark THE CORRECT choice.

Monoview/Multiview: Traditional tests report some aspects of the quality of your effort, translated into meaningless grade-level statements. Does anyone here know what "4.2 in math" or "equivalent to grade 10 reading" means? Our data set includes categories of performance-- correct, legible, requires improvement, learning opportunities, skips, to name a few--as well as presenting the rate of change through the family of Standard Celeration Charts. Changing the rate of change is our goal. We strive to maximize performance gains for each person.

Validity/Valid: tomes have been composed to justify the use of remotely chosen items presented by commercial tests in our performance settings. Enough said. (If you wish to study this topic from an historical perspective, check the history of "operational definitions" with a friendly psychologist. My, my!) Our data are valid, since when we monitor math, we measure our area of programming and of concern. This approach allows us to empirically verify our data, continuously, in each setting, on each project.

Metaphor/Relation: Perhaps, one day, it will be deemed immoral or unprofessional to translate raw data into an unknown? We do not know what age 2.6 on the Denver (or whatever test) means. We are unable to interpret what mental age 6.9 means. We cannot programme for a person who "scores" 8.2 on the language section of the CTBS, the ITBS or the FUTZ. On the other hand, directly quantified performance of specific topics, monitored over time aids everyone's understanding. We require clear awareness of relationships between events and performance. We've got it, let's use it.

If your head spinning? My suggestion is that you personalize these points. Play with the ideas a bit. If you don't need to dwell on the testing side, skip those points. concentrate on "how do we improve our monitoring?" That is the question.

Afterword

About 20 years ago, Og Lindsley presented ideas about the deficits of standardized testing (maybe in a course, perhaps at a local or national conference, maybe in a marathon rap session in some North American hotel room). He pointed out that we were in the process of standardizing the information format and flow relating to people and that we would gain significantly from our implementation of frequency monitoring along with Standard Celeration Charts. On the other hand, traditional testing worked strenuously to structure procedures-instructions, page format, administration minutiae as well as attempting to determine appropriate content, even sequences. Overconcern, and testing biases applied to inappropriate areas of classroom and research efforts contributes to weakening our people.

We regularly see performance levels seldom observed or recorded before our efforts. Our expectations are challenging. We support the behavers thoroughly, while delighting in their gains. We are humble in the realization of the magnitude of the task and of the potential gains to individuals and our communities associated with maximizing personal development.

Thank you for your attention. My next piece will explore the topic of data sensitivity. The

old terminology was Validity and Reliability, we are adding a crucial factor to our data concerns: Sensitivity.

RESOURCES

Gould, S. J. (1981). **The mismeasurement of man.** New York: W. W. Norton.

Haughton, E., (1972). Aims, growing and sharing. In Jordan & Robbin, Eds., **Let's try doing something else kind of thing.** Reston, VA: C.E.C.

Johnston, J. M., & Pennypacker, H. S. (1980). **Strategies and tactics of human behavioral research.** Hillsdale, NJ: L. Erlbaum and Assoc.

Journal of Precision Teaching. (1980-). McGreevy, P., Editor. Kansas City, MO: Plain English Publications.

Kunzelman, H. (1980). The learning metric. In **Human diversity and the assessment of intellectual development.**

Kunzelman, H., Ed. (1970). **Precision Teaching, an initial training sequence.** Seattle, WA: Special Child Publications.

Lindsley, O. R. (1964). Direct measurement and prosthesis of retarded behavior. **Journal of Education.**

Pennypacker, H. S., Koenig, C. H., & Lindsley, O. R. (1972). **Handbook of the Standard Behavior Chart.** Kansas City, MO: Precision Media.

Peters, T. J., & Waterman, Jr., R. H. (1982). **In search of excellence.** New York: Harper and Row Publishers.

Skinner, B. F. (1953). **Science and human behavior.** New York: Macmillan.

Skinner, B. F., Solomon, H., & Lindsley, O. R. (1954). A new method for the experimental analysis of psychiatric patients. **Journal of Nervous and Mental Diseases, 120,** 403-406.

Stevens, S. S. (1958). Measurement and man. **Science, 127,** 383-389.

Acknowledgements

Thank you for support and encouragement: Elizabeth Haughton, Claudette McGuire (graphics) and Diane Brownson (word processing).

COMPUTERS

Bill Wolking

AIMSTAR is an Apple II(+ or e) program for Precision Teachers. This column is a preliminary review of AIMSTAR. As far as we know, this is the first commercially available program for Precision Teachers. AIMSTAR is designed ". . . for the classroom teacher who collects performance data on her students but lacks the necessary time or support to carry out more formalized management procedures (e.g. graphing, plotting progress lines, or invoking decision rules)."

INSTALLATION. Insert the AIMSTAR disk into drive 1 and turn on your Apple. The program loads automatically. You supply the date and the number of disk drives you are using and the main menu is presented. Insert an initialized disk in drive 2 to store student data files. That's all there is to getting started. Unfortunately, there is no mention of how to make a backup copy of the master program disk. None of the copy programs available worked to make a backup. There is also no mention of how to replace a defective or lost disk.

PERFORMANCE. Eight menus control program functions: main menu, enter/edit student names, enter/edit student files, enter/edit data, chart data, execute decision rules, record strategy changes, and print reports. The menu system works well. Setting up a student file and entering data is straight forward. You are always given a chance to edit the data you just entered before it is saved. Starting a student file involves using the main menu and the first three menus listed on the main panel. First, enter the student's name and a three-digit identifier. Next, enter a description of the student's program—name it, classify as active or inactive (no longer inputting data), data type, frequency, duration, latency, percent), set the aim date, and the aim frequency. Then the option is presented to record what is called a textual description of the program. You may bypass this or use a mini word processor to write a brief description of: antecedent events, correct movement, incorrect movement, and consequence of correct and incorrect movements. Inputting chores are completed by moving to the next menu and entering the data for the student's program. A screen is provided for each data day. You record the date, number correct and incorrect, and number of minutes and seconds in the counting period. This screen also shows the number of the teaching strategy in effect.

Now you are ready to use AIMSTAR's four