A LEARNING METRIC

An Essay

By

Harold P. Kunzelmann, M.S. Ed.
International Management Systems, Inc.

April 1980

Requested for Publication

by

Mediax Association, Inc.
February 1980

in

Human Diversity and the

Assessment of Intellectual Development

A Learning Metric

I. Introduction

The setting is a typical first grade classroom. A pupil and examiner are seated comfortably in the rear of the classroom. The examiner says; "We are going to play a game. We are going to make loops. Like this (Demonstrates right side up to the pupil). You try right here. That's it - let me help. You've got it. Now let's play the game for real. You start making loops here. Are you ready? Please Start! (60 seconds later) Please Stop! Thank you - wasn't that fun? Please return to your seat."

The above is a simple example of a subtest being administered to a pupil for testing eye-hand coordination, beginning penmanship and direction following. Two different examiners (A and B) with different perspectives may do different things with the score the child obtained. These differences will be analyzed in detail to draw some important distinctions in the use and interpretation of pupil performance data.

Examiners A and B both score the loops the same. The criteria for a correct response is: looped up and crossed. The scoring by each examiner documents that the pupil made six (6) loops correct. The subtest scoring is reliable. The test score obtained by the child has face validity based on the relationship between performing loops and coordination, penmanship and direction following. The score of six (6) correct and none (0) incorrect falls in the developmental progression (i.e. generally younger children do less than six (6) and older children, 15 years + , do over 150 loops correct in a minute.). The two examiners next decide on what the pupil's score "means". That is they progress to the interpretative stage of performance analysis.

Examiners A: "A score of six correct loops is normal for a three year old, this pupil is six years and eight months of age. Table X of the manual states that the score (6 correct) for a six year old is equal to a mental age of 40 months. The pupil is 80 months old. The pupil is developing slowly on fine motor skills."

After matching all subtest scores to mental age equivalents the recommendations is, "The pupil suffers from developmental lag to the extent that mental retardation is suspected. Further testing and observations are necessary; however, at this time some special program is recommended."

Examiners B: "A score of six correct loops was a low performance today. I will repeat the entire test for 10 consecutive school days to determine if practice will produce improvement." For the next nine school days Examiner B repeats the test with the pupil.

At the end of 10 days the pupil scored 11 correct loops with one incorrect. At this time, Examiner B is ready to state what the score means. "With simple practice, basically repeated one minute measurements of writing loops, the pupil improves 35% a week with no instructional assistance. The overall improvement for the entire test was 20% a week having the lowest improvement on visual-verbal skills and the highest improvement on auditory-motor skills. The differential learning pattern of the pupil provides an optimistic outlook. Remedial action should include high performance aims for simple visual-motor and visual-verbal skills and an increased load (more complex) auditory-motor skills."

The two examiners' interpretation of the meaning of a score or overall tested output exemplify clear differences. Most notably Examiner A accepts one score, a snapshot, as an adequate base for analysis, interpretation and extrapolation. Examiner A, referenced the pupil's score to a norm which suggested an inferred fixed mental condition, that the pupil was possibly retarded. Examiner A concluded that the pupil needed remedial help with a limited chance to develop normally.

Examiners B's interpretation differed substantially from Examiner A's in that one score could only lead to a value or label whereas a series of scores could generate a refined perspective relating to improvement.

Examiner B could have found what Examiner A found. Examiner A could not have found what Examiner B found. The measures used by each examiner were the same, both are empirical. The parting point for the two examiners was what a score "meant". For Examiner A the score had value and metaphysical implication, whereas Examiner B pursued interventions designed to increase the pupil's capacity.

This paper offers an explanation and data supporting Examiners B's position, a position which focuses on the future rather than the past. A position which attempts to place human learning in a place of dynamic esteem rather than average conformity. A position which asks what human behavior Can Be rather than Should Be.

II. Premises

A. Learning is Improvement Toward Proficient Levels of Performance

This straightforward definition of learning avoids the massive theoretical constructs of what may or may not be occuring within a complex neurological system. This definition allows educational practitioners direct access to how their pupils are progressing or dynamically changing. Proficient levels of performance are being established.[1,2,3] Proficient performance levels in math, reading and social skills demand that average performance is only a benchmark toward excellence. For instructional purposes, average performance levels require additional teaching. The average or norm is not acceptable since learning is improvement toward proficiency.[4]

B. A Metric for Measuring Improvement is Frequency Over Successive Time Units

Frequency, a count of behavior for a fixed unit of time, is a natural measure of quantity and quality;[5,6] How much performance, performed how well defines speed and accuracy. Frequency measures in an instructional setting are efficient. The number of correct problems, correct words, correct concepts performed in a second, minute or hour may be measured by the learner and/or the instructor.

The statement, "Don't teach to the test," has no meaning when educational measurement is frequency. Summarizing frequency measures in relation to proficient performance levels negates the use of any test(ing). Instruction and measurement are interdependent not merely married.[7]

Historically, measures have been used to explore educational attainment. Binet and Cortis were early contributors to the development of relevant educational standards.[8,9] Binet's work has been prostituted to fit metaphysical notions about the brain's learning ability. Cortis's work involved standards for math skills based on one minute samples, resulting in the fundamental datum of frequency correct. However, historical surge toward the "perfect" test of mental capacities based on assumed validity has overshadowed his contribution.

More recently, the use of repeated frequency measures over successive intervals, has produced a Learning Metric. Lindsley, et.al. found improvement trends in day to day frequency measures.[10,11]

C. An Orderly and Sensitive Metric

Under fixed instructional assessment or testing conditions it has been found that the Learning Metric, correct counts/minute/week, is orderly.[12] Order may be defined as a systematic arrangement. Learning measures, the amounts of improvement, were low for some 10 to 15 percent of the pupils high for some 10 to 15 percent of the pupils and in-between for most. This orderliness exists across ages. Children, seven through twelve years of age, generally showed the same improvement distributions on skills taught at their age level. For a sample of over 8,000 children low learning (the 10th percentile score)

was 3% a week while high learning (the 90th percentile score) was 70% a week. Children at the 50th percentile, improvement was 30% a week.[13]

Improvement scores are sensitive to environmental conditions. Koenig et.al, found that generally more improvement occured for behaviors aimed at increased frequency when stimulus or instructional aspects of the environment were altered.[14] The reverse was true for decreasing behavior, where consequent or subsequent environmental alterations caused more improvement. These findings are based on over 30,000 behavior records each consisting of at least 15 days of frequency measures.

While establishing the orderliness and sensitivity of the Learning Metric, questions relating to what is being measured, how individuals and groups fair within the measured results and measurement consistency begged for answers.

D. The Learning Metric Sorts a Unique Aspect of Human Behavior

To add another and a more complicated measurement system to instruction and testing would be superstitious unless the measure improved education. Learning scores were correlated to I.Q. and achievement test scores. The finding showed no correlation (r= -.08 and -.12 respective).[15] The metric sorts a unique aspect of human behavior.

Secondly, repeated learning measures on the same skills, math and spelling, produced a test-retest correlation above +80. The metric is consistent.[16]

Last, the metric did not show any ethnic groups to be lower or slower than others. The improvement measures did not discriminate against minority children.[17] Only one of the eighteen significance tests performed showed any discrimination. If comparisons across ethnic groups have to be made for educational assistance or employment opportunities the minority person would be twelve times better off using the Learning Metric than either I.Q. or achievement measures.

The Learning Metric, frequency over successive time units, has order, is sensitive, measures a unique aspect of human behavior, is reliable and does not discriminate across ethnic groups.

E. The Learning Metric is Valid

Test makers, if they expect users to have confidence in the results of a test, must prove that the test indeed measures what they claim it measures.

This characteristic is most often called <u>validity</u> by test makers and measurement experts, and it is frequently expressed in terms of the degree to which one test correlates with another test already assumed or accepted to be valid. As one practitioner put it, "If a measure doesn't correlate to <u>something</u>, it is nothing." This practitioner's statement reflects the common view that validity is best established by correlating a new test to an older one, that is, by demonstrating so-called <u>concurrent validity</u>.

Such a view needs to be questioned, despite its long history of acceptance among psychological and educational testers. Why should one measure correlate with another? If one were to construct two identical buildings using different linear measures (inches and centimeters), one

would expect the two structures to be precisely identical. We do not need many tests to measure the same human development; rather we need one measure, to assess various components of the orderly processes of human development.

In the last several years an explicit goal of education has been the recognition of and attention to individual differences among learners. Tests are not always, if ever, expected to promote the development or discovery of identical learners. Rather, tests are expected to establish identical performance conditions for two different learners and to precisely identify each individual characteristic. If so, there is no justification for investigating concurrent validity between two tests.

More realistically concurrent validity is rooted in the competition between test makers to promote one test over another. To claim that a new test measures the same aspect of human behavior more efficiently than its predecessor is largely a strategy for marketing and selling a new test. In short, the current misuse of concurrent validity becomes a way of saying, "My test is the same as their's, but it's easier (or quicker) to give."

Test manuals often detail the studies establishing the concurrent validity with another similar test. But too seldom do those manuals focus on the relationship between what the test claims to measure and what performance the test actually measures. Herein lies the crux of any investigation of validity and herein users often find the weakest - or at least most highly inferential or hypothetical - of explanations. In contrast, note the behavior samples in our example at the outset of this essay. When a pupil had written what is indeed a loop, the question of validity is a simple one <u>provided the test claims to analyze and to predict loop-writing</u>. Whether any number of other tests, e.g., writing circles, writing crosses, will predict the pupil's loop writing performance is a moot question. The way to test loop-writing is to have the pupil write loops and measure this performance objectively and accurately.

In short, the simple counting and timing of directly observed behaviors is a valid measure of the speed and quality of human performance. And by repeatedly recording behavior frequencies periodically, the resulting Learning Metric is equally valid.

III. Directions and Observations

This essay explores the development and use of a Learning Metric. Frequency over successive time units (ie, counts/minute/week) defines the Metric which has met and stood up to most known measurement criteria and standards.

Continued investigations of the Learning Metric are taking the following directions:

A. What learning channels (stimulus-response modes) have the most impact in isolation or in combination, and in what sequence?
B. Do learning patterns change magnitude and vary more or less as a function of response calibration?
C. What effect does response variability have within and across learning patterns (improvements) for a person and across groups?
D. Are there critical performance frequency ranges necessary for future skill acquisition?
E. What proficient performances should be reached first, second etc. or is there a functional order?

Observations of applications of the Learning Metric suggest that its use will assist:

A. The practitioner who is not committed to labels but rather to future performance improvements.
B. The practitioner whose diagnoses always include proficient performance aims and never averages.
C. The practitioner who refuses to accept educational attainment which is measured by ethnically discriminatory devices.
D. The Professional Teacher and Administrative Associations that influence and initiate laws which demand that no test may be used to measure educational attainment which is not a direct part of the instructional process.
E. The pupil who wants to know:
    a. Where am I now?
    b. Where may I be?
    c. How will I learn best?

REFERENCES

1. Wood, S., Burke, L. Kunzelmann, H. and Koenig, C. Functional Criteria in Basic Math Proficiency. Journal of Special Education Technology, 1978, Volumn 2, 29-36.

2. Haughton, E. Aims-Growing and Sharing. In J.B. Jordan and L.S. Robbins (Eds), Let's Try Doing Something Else Kind of Thing. Arlington, Virginia: Council for Exceptional Children, 1972, 20-39.

3. Starlin, C.M. Evaluating and Teaching Reading to "Irregular" Kids. Iowa Dept. of Public Instruction, Des Moines, IA. 1979, December, 1-10

4. Haughton, E. Curriculum Options: Graduated Success to Better Mathematics, Special Education in Canada, 1973, 48, 14-21.

5. National Bureau of Standards: Time and Frequency,-(SP350); July 1955 - Dec. 1970; B.E. Blair, Supt. of Documents, U.S. Government.

6. Beck, R. and Clement, D. Precision Teaching in Review 1973-1976. Great Falls Public Schools, Great Falls, Montana, 1976.

7. Popham, J.W. Educational Measurements for the Improvement of Instruction, Phi Delta Kappan., 1980, April.

8. Cortis, S.A. Measurement of Classroom Students, New York General Education Board, New York 1919.

9. Binet, Alfred & Simon, Thomas. The Development of Intelligence in Children. L.C. 73-2962. (Classic in Psychology Ser.) Repr. of 1916 ed. (ISBN 0-405-05135-2) Arno.

10. Lindsley, O.R. From Skinner to Presicion Teaching: The Child Knows Best. In June B. Jordan and Lynn S. Robbin (Eds.) Let's Try Doing Something Else Kind of Thing: Behavior Principles and the Exceptional Child. Arlington, Virginia: Council for Exceptional Children, 1972.

11. Penneypacker, H.S. et.al. Handbook of the Standard Behavior Chart, Kansas City, KS., Prec ion Media, 1972.

12. Kunzelmann, H.P. (Ed) Precision Teaching: An Initial Training Sequence. Seattle WA.: Special Child Publications, 1970.

13. Koenig, C.H. and Kunzelmann, H.P., Learning Screening Manual; Charles Merril Publishing Co., Columbus, OH. 1980.

14. Koenig, C.H. The Behavior Bank: A System for Sharing Precise Information. Teaching Exceptional Children , Spring 1971, 157.

15. Kunzelmann, H.P. and Koenig, C.H., REFER Screening Manual, Charles E. Merril Publishing Co., Columbus, OH. 1980.

16. Caulfield, Hal, Disseration in Progress, Dept. of Special Education, University of Washington, Seattle, WA., 1980

17. Progress Report II: Learning Screening, State of Washington, 1976.