READING RESEARCH: STUDIES AND APPLICATIONS

Twenty-Eighth Yearbook
of
The National Reading Conference

Edited by
MICHAEL L. KAMIL
and
ALDEN J. MOE
Purdue University

with the editorial
assistance of
CAROL A. DAVIS
and
R. TIMOTHY RUSH
Purdue University

Published by
The National Reading Conference, Inc.

1979

J. JAAP TUINMAN
Simon Fraser University

BEYOND CRITERION-REFERENCED MEASUREMENT

Presidential Address presented at The Annual Meeting of The National Reading Conference, St. Petersburg, Florida, November 1978.

The other night, traveling from here [St. Petersburg] to Orlando I passed a sign announcing a religious event. In bold neon letters the sign read:

REVIVAL — Reverend Harry Dull

You laugh and I of course anticipated that. The concepts of dullness and revival are incongruous; their being contrasted, presumably unintentional, we find amusing. As language users we have no difficulty identifying and "explaining" this and similar linguistic accidents. As psychologists of language, however, we are yet less accomplished. Are our theories of language, of implication, of behavior, both general and specific enough so that we can build a machine that would chuckle when presented with this neon sign? I doubt it. And, I'm sure, our colleagues in artificial intelligence would be equally skeptical. Nevertheless, real progress in the measurement of reading skills depends intimately on the further development of an adequate psychology of language. My remarks this morning should be interpreted with this proviso in mind.

This is the year Oscar Buros died. Those interested in the content of educational tests lost a friend. Throughout his career Buros emphasized that no amount of statistical manipulation can remedy the weakness of a poorly conceptualized test. His vigilance against uncritical acceptance of psychometric evidence as proof of the quality and usefulness of educational tests I believe unmatched.[1]

Educational tests, reading tests not excepted, suffer far more from meager mathematical justification. This is an old complaint. In her classical article on item analysis Marion Richardson noted even in 1936:

> "The present writer is of the opinion that the ingenuity displayed in the invention of new indices has outstripped the critical examination of the logical foundation for item analysis." (Richardson, 1936, p. 395)

Forty years later, in an insightful review of key developments in educational measurement in this decade Lumsden (1976) addresses the same issue from a slightly different perspective when he calls for a new kind of test theorist:

> "They will not test a new model with a few items from the SAT files (or from a computer), find a mediocre fit to some dubiously relevant criterion and then to on to the next. Rather, they will set out the requirements, testing the model against a user-for-blood standard of efficiency. They will not seek salvation in the epicene elegance of elevated algebra but will prefer vulgar analogies" (p. 277).

Buros, Richardson, and Lumsden make an identical point which cannot be overemphasized. If educational measurement is in an impasse, and measurement of reading skills *is,* the problem is not with the algebra, but with the thought behind it.

Some four years ago I introduced an analogy to illustrate some basic conceptual problems facing criterion referenced measurement (Tuinman, 1978). Briefly, the story goes as follows:

"Once upon a time, in the chilly, icy, most southern regions of the globe, a tribe of Penguins decided upon an extensive self-improvement program. A committee of new experts in the art of self-improvement met and decided on the goal of the first Federally Funded SI Program. 'We must,' they said, 'learn to fly.' After reading a book by a learned Penguin named Pipham, the committee realized that their global goal needed to be broken down into narrower objectives. 'Let's make them behavioral," a farsighted Penguin suggested. 'That way we will have fewer measurement problems and less uncertainty about the success of our training programs.'

"Quickly it became evident that, in order to succeed in formulating adequate objectives, the behavior 'flying' needed to be observed first hand. So a subcommittee was commissioned to trek north and observe flying animals in different situations.

"One member of the committee made a special study of seagulls. He noted that every time just before these birds soared into the air, they dipped their heads under water. Dutifully he made note of this very important characteristic.

"One of his colleagues reported that the birds he observed would peck into the ground and extract a worm just before taking off. The committee decided that the two behaviors showed a great deal of similarity and that, no doubt, both contributed to the mechanics of taking off.

"During a month of study, many more analyses were made of the art of flying. The discovery which pleased the committee most, however, occurred when they inadvertently stumbled upon Dallas Airbase and observed the take-off of truly big and heavy birds. 'We must analyze their behavior very closely,' they told themselves. 'Those are the only birds bigger than us.' Thus, they added to their list of objectives:

01. 'The student must be able to run at great speed in a straight line before taking off.'
02. 'The student must be able to produce a loud whirring sound from deep inside his/her chest before starting the take-off run.'

Satisfied, the Subcommittee on Objectives returned to Penguin Land.

"Soon a training program was implemented and criterion referenced measures from each objective were built and validated.

"After three months, the first student (a smart Penguin cookie as ever there was) had completed all training modules and passed all criterion referenced tests. She was ready to fly. The Committee on Self-Improvement came out in full force to observe and celebrate the event.

"The student's whirring sounded beautiful, her pecking at the hard ice was rhythmic and smooth, her running down the take-off strip an awesome display of athletic prowess. Alas, she never took to the air, and after a full day of frenzied flying was finally committed to a hospital to recoup from a nervous breakdown."

At that time, my interpretation focused foremost on the license to build tests endlessly suggested by the task analysis performed by the birds. Now, I'd like to emphasize the non-theoretical nature of this type of task analysis as at least a partial cause of the birds' unjustified optimism. Parenthetically, I should also point out that I am now far less sanquine about the prospect of constructing

meaningful and useful learning hierarchies of reading skills than I was when I wrote the story. In general, it appears to me that criterion referenced measurement in reading has made little or no progress after the initial application of CRM thought to our area of curriculum specialization.

Today I want to specifically emphasize three points in discussing the measurement of reading achievement:

a) Norm referenced measurement needs shoring and re-interpretation, but it is here to stay.
b) Technical problems with criterion-referenced measurement will be resolved only if and when certain conceptual ones are dealt with first.
c) Our focus should be on educational measurement rather than on educational tests.

Let me be blunt and state that I consider norm-referenced measurement both unavoidable and desirable.

The human need to integrate, to interpret information to the largest possible unit is illustrated in our sphere by at least two measurement "events." First, after the initial reading assessment, NAEP actively sought the cooperation of the IRA to establish an interpretation panel, to "make sense" out of the data. In fact, the scores on each of the criterion items are fairly unambiquous. Problematic is the relationship of all these separate pieces of information to some less atomistic unit. Second, consider the recent spate of papers on the RMC models presented at the 1978 AERA meeting. (Fishbein, 1978; Linn 1978; Barnes and Ginsburg, 1978; Rutherford, 1978). As a group these papers present the spectacle of statisticians scurrying to retrieve meaning. In this case the move is not only toward larger units of interpretation but as well toward an undisguised reinterpretation of CRM data in NRM terms.

The human need to compare is so dominant that, I fear, no amount of psuedo-philosophical theorizing about this form of measurement in education is going to thwart its expression. In many contexts the value of the achievement is in the comparison.

When the 1978 Nobel prize for Peace was awarded to Mr. Begin as well as to Mr. Sadat, Time Magazine writers made an explicit comment on the fact that the prize lost in meaning for Sadat *because* Begin also was included. In sharing the achievement dulls. The single book which may represent a towering achievement in one faculty will be of far less merit in another faster-publishing group of academicians. The four minute mile long was the target of many a runner. Many may have thought of it as an absolute criterion. However, when Bannister broke the magic limit his performance was important only because no one else had done so. There is no doubt in these and many other instances that the value of an achievement is determined and determinable by comparison.

In practice the fine tuning of criterion-referenced tests in fact leads to a kind of hidden norm-referencing. When, as is customary, items are tried out, their retention in or rejection from the final test often depends on how difficult they are for the students in the pilot study. The performance of a new group of students, therefore, depends in large measure, on the level of skill of the students in the initial pilot. The statement "The new students scored at

mastery level" needs to be complemented: "on a criterion suitable for students of X level of skill." Criterion-referenced measurement of reading skills than becomes in effect indirectly norm-referenced. Johnny only is a master when compared to thirty specific other Johnnies. Given thirty Sams, of higher ability, in the pilot, Johnny might not have been able to reach mastery level.

Finally, reading tests involve the interaction of mental operations and linguistic elements. The trouble is that the operations, themselves varied and complex, are performed on essentially an infinite number of combinations of an extremely extensive set of linguistic elements. The ability "to draw casual inferences" can be tested in such a variety of syntactic, lexical, semantic and pragmatic contexts that it becomes virtually impossible to specify meaningful, non trivial item writing rules which operationally define criterion behavior. (I recognize the work of Bormuth (1970) and others but I have, among others, serious reservations about the practical contribution of that kind of suggestions (Tuinman, 1970, 1978).) The best the writer of criterion-referenced measures can hope for is a universe firmly defined by a specific curriculum, textbook, story, etc. from which to sample. The application of CRM is more generally defined contexts is severely limited, however, in plain language: There just are too many loose ends regarding the texts to test on; the information in these texts to focus on: the question to ask; the way to ask the questions. None of these issues are terribly problematic in a theory of norm-referenced measurement. In CRM, however, they represent severe problems in terms of identifying criterion behaviors and in terms of selecting intrinsically definable scores.

We might benefit from a change in terminology. "Norm-referencing" has connotations which are unhelpful. Typically, norm-referenced tests assume that the ability of achievement measured is normally distributed in the population. Hence, tests are constructed in order to show normal distribution. I need not argue here that this assumption may have little merit for the measurement of at least some aspects of reading achievement. (See also, Lumsden's (1976) argument against assuming any "underlying" or "true" ability.) Comparison of one individual's performance to that of an appropriate norm group is possible even if the tests used do not assume or force normal distributions. We need comparative interpretation tests based on meaningful performance curves. I view such curves as *de facto* descriptions of the performance of appropriate norm groups. Their construction requires a pulling-up-by-the-bootstrap-procedure, initiated by the use of naive items (i.e. *not* preselected on an assumed mathematical model). In addition to being naive, these items must be *realistic*. That is, they must be derived from descriptions of actual usage of information. Comprehension questions, for instance, should stem from analysis of the internal question asking process the reader engages in as he reads. Taxonomies are fine for defining the range of intellectual operations to be engaged in; they say little about their frequency and context. As such they form an inadequate basis for measurement.

Comprative interpretation tests require the establishment of meaningful reference groups. Age and sex in many instances are only of apparent interest because they are traditional. How well does Johnny read compared to other science students, compared to middle level bureaucrats? The use of such

comparison groups assumes resolution of the difference between using the average engineer and the engineer's average as a norm.

Comparative interpretation tests, under ideal circumstances define performance both in terms of a what's read and who else is reading it; Y reads X-material as well as Z-persons.

The problematic nature of CRM is illustrated in much of the current work on technical concerns. Illustrative is a recent paper by Berk (1978).

Specifying a 95 percent confidence interval he is able to show that one needs 58 items to test the comprehension of a population of 1000 sentences if one uses 80 percent mastery as the criterion score. This sounds very precise and helpful until one realizes that Berk's calculations are based upon Bormuth's unproductive conceptualization of the measurement of comprehension (Bormuth, 1970; Mehrens, 1970; Tuinman, 1970).

That the development of new indices does not necessarily even reflect progress on the technical side is demonstrated by Downing and Mehrens (1978). When these authors compared six single-administration reliability coefficients for CRT's, they found only one measured test characteristic that differed from the classical Kuder-Richardson formula's. (This study did not include Brennan and Kane's (1977) index. However, see Lumsden (1976) for a critical evaluation of the signal/noise ratio concept.) More disturbingly, when Smith (1978) compared five popular item selection methods he found that none of the methods was consistently superior. Indeed, random selection of items worked just about as well as any other method.

The general point that the elaboration of mathematical techniques can obscure basic conceptual problems is yet more clearly demonstrated when one reviews recent work on establishing adequate criterion, mastery or pass scores. The intent, of course, is to formulate techniques which will validly classify individuals as masters or non-masters. Two of the more interesting papers by Huynh (1977) and Faggen (1978) illustrate current attempts to deal with the issue and their inadequacy.

The procedures for establishing mastery scores are elegant and appear effective. Nevertheless, both papers share a very fundamental problem. Whether or not a student is validly classified as a master on test i is judged by his/her performance on test j. Presumably, test j is more complex, more general or higher in some skills hierarchy than test i. I grant that adequate tests of accuracy of classification are needed. The bulk of the unresolved issues, however, relate precisely to the specification of the relationship between tests i and j and to the adequacy of test j as a criterion. This is far from an issue for idle contemplation. It raises its head in very reading management system currently in use. Test constructors are saved by the fact that most tests, whatever their label, format and content, load heavily on a general verbal factor. Were it not for that fact, teachers would be confronted with many more puzzling patterns of performance on series of tests in their "management" batteries than they already face.

Seldom are relationships among tests made explicit. Implicitly, however, hierarchical connections and transfer of skills acquired to other more complex ones are routinely assumed. Criterion-referenced comprehension tests fail, however, precisely because either they can't satisfy these assumptions or, in

the event they do, there are precious few ways to empirically demonstrate this.

An early optimism about the possibility of isolating and describing learning hierarchies is now tempered by the reailization that even in the case of simple behaviors, empirical verification of hierarchical relationships is difficult. Two recent papers demonstrate the paucity of effective techniques. Guay and McCabe (1978) point out serious shortcomings in existing tests of hierarchical dependency. They then present a test which they claim remedies these problems. However, being restricted to a pair-wise comparison of skills their test too is very limited in scope. The state of the art is best illustrated by Griffith and Cornish (1978). They extend White and Clark's Test of Inclusion to more than three items per skill. The relationships among 7 skills, representing the development of an introductory chemistry concept, were than analyzed. In all, some 10 possible hierarchies result from the analysis. Hardly a comforting thought, considering the relatively simple nature of the skills involved. The prospect of extracting ourselves from the measurement muddle through specifying skills hierarchies in the comprehension domain is dim indeed.

Even if more powerful statistical techniques were available and practical, there is still grave doubt regarding the applicability of the hierarchy concept to the comprehension domain. Every teacher knows that it is possible to ask a very easy "higher order" question and a very difficult "lower order" question. If skills hierarchies are to be studied at all, this should be done within a given level of language, within a defined lexicon. We will never understand the contribution of "being able to use content" to "being able to draw inferences" unless we first systematically study the relationships among such skills for a particular set of words, specific syntactic patterns, and a specific semantic and pragmatic context.

No doubt CRT's find their justification in their potential utility for making instructional decisions. Yet, in practice, teachers as often as not find current CRT's unhelpful or confusing. The relationship between specific lower level tests and more global achievement is often tenuous and counterintuitive. The problem lies only in part in inadequate validation of learning hierarchies. Another shortcoming of CRT's of the type currently in vogue, however, is their exclusive focus on product at the neglect of attention to process.

Cognitive psychology emphasizes an orientation towards the learner who monitors his cognitive processes according to task demands. In comprehension the emphasis on active processing is illustrated in the work of Wittrock and his associates (e.g. Doctorow *et al.*). Formally, Greeno (1976) proposes a reconceptualization of learning objectives, underlying CRM, in terms of outcomes and cognitive process analysis. Instructional measurement must be reevaluated in terms of the cognitive demands of both criterion tasks and enabling skills. Such a reevaluation is a condition *sine qua non* for progress in the development of instructionally useful criterion-referenced measurement.

Considerations of validity and reliability are typically limited to a test *per se.* Yet, tests differ widely in their potential for misuse and in their robustness against misinterpretations. In the end, in instructional contexts, most tests are used to assign pupils to a limited number of nominal or ordinal categories. I would like to see our definitions of reliability and validity to include that final step in the measurement process. Hence, for instance, a reliable test is one which results in a user consistently making the same instructional decisions; or alternately, a test which leads different users, given a set of prescribed instructional options, to make comparable instructional decisions. Whether or not this proposal is too radical or too cumbersome, I believe that we sorely need studies of the relative amount of variance due to users and due to tests in situations where tests purportedly form the basis for instructional decisions.

There is no use for a ruler which permits measurement to the millimeter when there is no one who can hold it without a tremor.

REFERENCES

Barnes, R.E. and Ginsburg, A.H. *The relevance of the ROM models for title I policy concerns.* Paper presented at the Annual AERA Meeting, Toronto, 1978.

Berk, R.A. *Item sampling from finite domains of written discourse.* Paper presented at the Annual Meeting of the AERA, Toronto, 1978.

Bormuth, J.R. *On the theory of achievement test items.* Chicago; University of Chicago Press, 1970.

Brennan, R.L. and Kane, M.T. Signal/noise ratios for domain referenced tests. *Psychometrika,* 1977, *42,* 609-625; 1978, *43,* 289.

Doctorow, M., Wittrock, M.C., and Marks, C. Generative processes in reading comprehension, *Journal of Educational Psychology,* 1978, *70,* 109-118.

Downing, S. and Mehrens, W.A. *Six single-administration reliability coefficients for criterion-referenced tests: A comparative study.* Paper presented at the Annual Meeting of the AERA, Toronto, 1978.

Faggen, J. *Decision reliability and classification validity for decision oriented criterion referenced test.* Paper presented at the Annual Meeting of the AERA, Toronto, 1978.

Fishbein, R.H. *The use of nonnormed tests in the ESEA Title I evaluation and reporting systems: Some technical and policy issues.* Paper presented at the Annual AERA Meeting, Toronto, March 1978.

Greeno, J.G. Cognitive objective of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), *Cognition and instruction,* Hillsdale, N.J.: Erlbaum, 1976.

Griffiths, A.K. and Cornish, A.G. *An analysis of three recent methods for the identification and validation of learning hierarchies.* Paper presented at the Annual Meeting of the AERA, Toronto, 1978.

Guay, R.B. and McCabe, G.P. *A chi-square test for hierarchical dependency.* Paper presented at the Annual Meeting of the AERA, Toronto, 1978.

Juynh, H. Two simple classes of mastery scores based on the beta-binomial model, *Psychometrika,* 1977, *42,* 601-608.

Linn, R.L. *The validity of inferences based on the proposed Title I evaluation models.* Paper presented at the Annual Meeting of the Aera, Toronto, 1978.

Lumsden, J. Test theory, *Annual Review of Psychology,* 1976, 251-280.

Mehrens, W.A. Scientific test construction — pure and sterile. *Contemporary Psychology,* 1970, *15,* 666-67.

Richardson, M.W. Notes on the rationale for item analysis. In W.A. Mehrens and R.L. Ebel (Eds.), *Principles of educational and psychological measurement.* Chicago: Rand McNally, 1967.

Rutherford, W.L. *Implementing criterion-referenced instructional programs: truth or fiction.* Paper presented at AERA, Toronto, March 1978.

Smith, D.U. *The effects of various item selection methods on the classification consistency of criterion-referenced instruments.* Paper presentation at the Annual Meeting of the AERA, Toronto, 1978.

Tuinman, J.J. *Bormuth's views on the theory of achievement test items — some questioning comments.* Paper presented at the Annual Meeting of the National Reading Conference, St. Petersburg, Florida, 1970.

Tuinman, J.J. Experimental research in reading — some design considerations. In W.E. Blanton & J.J. Tuinman (Eds.), *Reading: Process and pedagogy.* Viewpoints, 1972, *48,* 99-107.

---

[1]I permit myself a personal anecdote to illustrate this. After reading the first draft of my review of the Woodcock Reading Marking Tests for the Eighth MMY, Dr. Buros commented that he liked my emphasis on item content but lamented my rather unquestioning acceptance of the elaborate and unique statistical apparatus accompanying the test. The subsequent revision of my manuscript incorporated many of his incisive critical comments. I assume that, in this way, Buros exerted a healthy influence on test reviews, and test reviewers, throughout his career as MMY editor.

J. JAAP TUINMAN
Simon Fraser University

AN INSTRUCTIONAL PERSPECTIVE ON BASIC RESEARCH IN READING

Presidential Address presented at the Annual Meeting of The National Reading Conference, New Orleans, November 1977.

My major contention today is that reading researchers do not spend enough time in the schools and classrooms when reading is taught and learned. More generally, reading researchers have too little contact with the reader as he learns and reads.

I shall analyze some of the reasons why this is so and speculate on the consequences.

A survey of the research report in such journals as the *Journal of Reading Behavior,* the *Reading Research Quarterly,* the *Journal of Verbal Learning and Verbal Behavior* reveals that (a) the main reason for researchers to be with readers is to collect data from them and (b) this data collection process is often very brief. It is not unusual for a reading researcher to collect all the necessary data for a study in 3 to 5 hours.

The fact that reading researchers are in contact with readers only at the moment of data collection is far from trivial. The history of science is filled with accidental discoveries and insights. Quite likely it is superflous to refer to Galvani's discovery of the principle of the electric battery after he watched a frog's leg twitch and to Minkowski's discovery that urine contained sugar (an important discovery for medical help of diabetics) when he was operating on his dog. I believe that the *modus operandi* of many reading researchers minimizes the opportunities for such chance discoveries.

Reading researchers have no laboratory or they behave as if they don't. The *a priori* specification of the kinds of data to be collected, the quick in and out data hunting forays into the classroom minimize the opportunity for the fortuitous accident and, more importantly, for the creative generation of valuable hypotheses.

The classroom, the reading corner, the library are the reading researchers' laboratory. Most of these stay vacant.

A scientist must develop a sense of his phenomena and a feel for his data. The latter aspect is stressed frequently by means of such exhortations as: "plot your data before you try to interpret your correlations, etc." Little emphasis is given, in contrast, to opportunities to develop a sense of the aspect of reading studied other than in the forms of scores, data. Yet, such sensitivity is essential. As Flesch (1951) points out Roentgen discovered X-rays by accident when he noted that cathode rays penetrated black paper only because he was Roentgen, sensitive to the data coming his way which others might have thought meaningless. There is nothing mystical about this kind of insight. More often than not it presupposes intimate experience with the phenomena under study.

Our experimental designs and even more our research traditions do not encourage long term study of children's reading in actual learning situations. In this respect the North American researchers tend to differ from their West European and Russian colleagues. I refer to the work of Peter and Else Petersen (1965) and to the studies of Gal'perin (1972).

Part of our problem is an all-consuming pre-occupation with outcomes of the learning process rather than with the structure of the process. I suspect that if the development of psychometrics during the last 75 years (and particularly during the first half of the period) had been less rapid, the study of both the reading process and the teaching of reading would have taken a quite different turn.

The Dutch educational psychologist Van Parreren (1970) maintains that psychologists and educators (he uses the term didaxologist, one studying and learning and teaching) take a different view of learning. The psychologist normally limits himself to the study of variables affecting the learning outcome. One assumes the end results of a learning process the crucial measure of that process. The educational psychologist, by contrast, Van Parreren says should be primarily interested in those actions, behaviors of the learner which result in changes of performance. The task of the educationally oriented psychologist becomes fourfold:

a) analysis of performance results
b) analysis of the states the learner traverses to achieve these results
c) an analysis of the conditions of the learning process
d) an analysis of the future functioning of the learning achieved.

Though these insights may not be startlingly novel they place an emphasis on close and enduring contact between researcher and reader which, I maintain, is not now a part of our research tradition.

The contrast sketched above is sometimes couched in terms of quantitative analysis and qualitative analysis of behavior. It strikes me that an increase in the frequency of observations on readers may go a long way to detecting and documenting qualitative shifts in performance. (I deliberately duck the philosophical issues involved in reducing quality to quantity). A simple example may clarify what I have in mind. In studying the acquisition of synonyms I teach a set of synonym pairs over a two week period. At the end of the treatment I measure each student's performance and relate it to such variables as verbal fluency, reinforcement schedules, or what have you.

What I don't know is that three of the eight experimental subjects for the first four days learned by rote association the word pairs involved. Only then did they realize (discover, etc.) the principle of "same as" involved.

I maintain that this kind of information is important and, moreover, that reading research has done a poor job of zeroing in on these qualitative changes in the learning process. The branch of psycholinguistics associated with the Goodmans employs a methodology conducive to the kind of research I suggest, but their range of phenomena is too restricted and their isolation from traditional hypothesis testing too absolute.

The work of Hansen and Lovitt (1977) is another illustration of the commitment to the kind of long term, intensive data collection I have in mind. They use *Applied Behavior Analysis* to teach subjects' acquisition of basic reading skills, mostly decoding. ABA, these authors say, is characterized by direct and frequent measurement, analysis of individual data and experimental control.

ABA comes from a very narrow tradition of behaviorism and has found distinct favor with learning disability researchers. This, I think, is incidental to the focus of ABA. Its contribution lies in the concept of intensive monitoring of

the teaching/learning process and the admitted relevance of data on individual subjects.

I am not yet arguing that we can build sound theories on data obtained from small sets of students but I do maintain that (a) the source of reading research lies in constant observation of the reading and (b) that the purpose of predicting future behavior data from a limited set of real individuals may be as useful, or more so, than data on the mythical average individual produced by statistical analysis. Prediction by analogy (Johnny is like Mary, so,. . .) can be safer than prediction by generality. From an epistemological view this position is far less naive than it may appear at first.

One way of dealing with complex phenomena (such as the reading of these children, in this book, in this classroom, etc.) is dealing with them in as concrete a fashion as is possible. Hence the analogy approach touched on above.

A quite different approach to complex phenomena is through statistical decomposition and reconstruction by identifying simple variables making up the complex and by relating them in some kind of mathematical model. Fisk (1977) makes a strong argument from a philosophical point against the assumption that we can know complexes from their constituent parts. In the psychologists' language his position is Neo-Gestaltist: not only is the whole more than their parts, the function of the parts can only be known in their relationships to the whole.

To make this more concrete, I note that in the October issue of the *Journal of Verbal Learning and Verbal Behavior,* Levy (1977) opts for an interactive model of reading in order to explain the relationship between speech (or sound) analyses of printed text by the reader. She also notes that the "relationship is not straight-forward". I can imagine numerous adoptive relationships between different levels of decoding and language skills and different text demands in terms of decodability, linguistic and cognitive characteristics of the text involved. I see no point in studying the contribution of decoding to comprehension without taking in account the states of each of the other variables (and I only mentioned a subset of the cognitive variables involved).

Kerlinger in his recent AERA presidential address (1977) spent some time illustrating that methodology has a profound effect on practice by pointing to the emergent technique of analysis of covariance structures. I quote: "It integrates factor analysis, including hypothesis-testing factor analysis, multivariate analysis, study of change and path analysis for example, in a framework explicitly oriented to theory and hypothesis testing." (p. 9)

We are all very familiar with the limitations of analysis of variance *vis a vis* complex phenomena. The partialing out of more and more variables, a process which should be based on theoretical considerations but hardly ever is, is severely hampered by the difficulties of interpreting and representing higher order interactions. Any technique which can handle complex events more adequately is welcome.

Yet, I wish to note that statistics should remain a last resort tool of the social science researcher. Powerful statistics have hidden much of the dubious nature of the sociological laws derived from questionnaire data, for instance (Borgatta, 1969). Of all people Oscar Buros (1977) holds that advances in

statistical sophistication hindered conceptual progress in test construction. Again, in the inaugural number of *Instructional Science,* now five years ago, Meredith (1972) notes:

". . .we need to ask more and more simple questions if research is to cohere. Otherwise we hand ourselves over to the statisticians and this means letting their mathematical structure determine the structure of our knowledge. . .which is precisely what factorial psychology has done. It represents an abdication of thinking and a denial of sense in the original test data. Any record of actual performances has a good deal of differential sense in it, but once we start scoring, summing, calculating means and standard deviations and correlations we have abandoned sense. . ."

In the view of Meredith (a student of Spearman incidentally) the first obligation of a researcher is to feel the sense of this phenomena. Therefore a student of behavior must get a sense of behavior. To return to my theme: reading researchers should be in their classrooms or very near them.

For whatever reasons we have created a very false and disturbing climate for reading research in our schools of Education. (Parenthetically, even a basic research propagandist as Suppes places the responsibility for educational research squarely within facilities of education, (Suppes, 1974)).

A number of years ago I established that the education faculty of one particular major university in the U.S. on the average published less than one publication in a refereed journal every two years and, more disturbingly that many faculty members hardly published at all. Recently Arlin (1977) using the same methodology but on a much grander scale collected evidence of lack of sustained inquiry among educational researchers across North America.

In an analysis of eight years of professional journals abstracted in the *Current Index to Journals in Education,* Arlin found that (a) about half of all publications were produced by one-time authors and (b) that between 60 and 75% of all the authors associated with the 130,000 plus articles analysed, published only once during the eight year period studied.

Arlin laments the "lack of sustained inquiry", and rightly so. I suggest that the situation in effect is more perturbing than his data reveal. The writing of many multiple article authors, a proper analysis would reveal, is characterized by frequent shifts from topic to unrelated topic.

We are all familiar, perhaps too painfully so, with the faculty member strong in research methodology who today publishes an analysis of the acquisition of "French speech patterns" and tomorrow an account of "educationally instructive initiation rites among Canadian native Indians." Someone once suggested an index of scholarly contribution (SC) as follows:

SC = 1 - A/p, where A/ = the number of different areas of inquiry engaged in and

p = number of publications

I suggest that for many reading researchers this index would approach zero.

I am not harsh on individuals. I do, however, fault the currently prevailing training and reward systems in major universities. Conditions for which, I might add, we share considerable responsibility.

As to the training of reading researchers within faculties of education too little emphasis is placed on historical and conceptual analysis of the phenomenon of interest. The term "research skill" carries the connotation of

"data analysis" too frequently. Philosophy of science and practical exercises in critical and creative thought (dare I suggest reading research variations of that favorite parlor game 20 questions?) routinely take the back seat to "introductory statistics" 1, 2, and perhaps 3.

I am by no means suggesting a reduction in exposure to techniques of design and analysis. I am saying that in most programs of training for scientific research the stress is on research at the expense of the requirement that this research be scientific.

A moment ago I mentioned existing professional reward systems. I am convinced that the relative absence of continuous and sustained research among productive members of faculties of education is in part due to the pressure to produce quantity rather than quality of work.

Many of us do research for professional rewards. Perhaps someone will produce a study sometime validating this assertion by showing peaks of production in periods immediately prior to promotion and tenure decisions. The criteria for rewards have, whether we like it or not, a definite influence on the research production. The lackadaisical researcher receives periodical impulses to churn out some indifferent work. Far more serious are the facts, however:

a) that many young and capable faculty members believe it to be against their best interests to give themselves time to pursue the study of a particular phenomenon in leisure and with thoroughness and
b) That they, rewarded in measure, come to believe to have embarked upon a career of scientific research.

Improvement of reading research, the strengthening of a genuinely scientific study of both the learning and, of the teaching of reading requires patient manipulating of non-trivial variables in what I refer to as a classroom or a real learning context and what Uri Bronfenbrenner (1976) designates as the learner's micro-environment. Such study requires thorough knowledge of the history of the variables manipulated and of their epistemological status.

Many of the persons in this audience are senior members of their respective faculties, often involved in its administration. My appeal to you is to increase the emphasis on the quality of inquiry in assigning professional rewards to colleagues.

Reading Departments and departments of Educational Psychology have demonstrated adequate competency in doing studies. For many the time has come to start doing research.

Kerlinger, in the AERA address I referred to above, has made a strong plea for support of basic research in education. The arguments he advanced are straightforward. Dominant is his belief that basic research has more potential for practical pay-off than many applied research efforts. Much of the support for this belief is drawn from the Comroe and Dripps (1976) study of advances in medical practices. Kerlinger notes that basic research was responsible for almost twice as many "key articles" than non-basic R and D taken together.

I have no desire to debate Kerlinger's position in general. However, he addresses himself to reading specifically, in a manner immediately relevant to my topic today. Permit me a rather lengthy quote:

"Answers to reading problems lie not in many researches aimed at telling teachers how to teach reading. They lie in research aimed at understanding

the many aspects of human learning and teaching connected with reading. . .Study of reading in and of itself is almost invariably unproductive. We must study reading in the context of perception, motivation, attitudes, intelligence and so on. In other words the goal should not be the improvement of reading. It should be understanding of the relationship among the many complex phenomena related to reading. Research directed to improving anything but minor skills is doomed to triviality, frustration and defeat. To improve something as complex as reading requires understanding of reading and many related phenomena, a very difficult task indeed. And there is, of course, no guarantee of improvement in children's reading, even if basic research on phenomena related to reading is done." (p.7)

I understand and, in part, sympathize with Kerlinger's rejection for pragmatic practical pay-off and for relevance. It is clear that his tenure at the University of Amsterdam with its strongly Marxist oriented Education faculty causes him much intellectual discomfort and may have strengthened his resolve to safeguard disinterested basic research. I do have a number of critical observations on his perspective, nevertheless.

First, it is not clear to which degree Kerlinger distinguishes between "research aimed at telling teachers how to teach" and basic inquiry in how people do teaching. Surely we must admit of basic study of the variables involved in teaching reading and their relationships.

Secondly, from his examples in this quote and from comments throughout his paper it is clear that Kerlinger insists on equating basic research with foundational or disciplinary research.

In this he is in the company of such distinguished COBRE (Committee on Basic Research in Education) propagandists as Carroll and Suppes (1974) who identify no less than 12 disciplines related to education ranging from psychology to the biosciences.

Where as, no doubt, reading research can benefit from advances in many of these related fields there are no *a priori* reasons why inquiry in educational processes must be limited by the epistemology embedded in the "foundational" disciplines. This needs repeated stressing.

Psychologists, as we know them, will never understand how children learn to read. The reason is simple. Most children learn in an environment in which many factors interact continuously. Moreover children have histories the influence of which may be cancelled out or minimized in psychological experiments but which most definitely will play in the actual learning process. Learning to read, learning to comprehend better, in schools is an educational process, not merely psychological, or sociological, or biological or linguistic. A confluence of perspectives at one time viewing one phenomenon is needed.

Western scientific thought pushes for identification of simple and isolated variables. An attending danger is that we may destroy the phenomenon we wish to study and produce irrelevant basic research. (That is, basic research on pseudo phenomena). In surveying the many (quasi) psychological, (quasi) sociological, quasi etc. treatments of "reading" which surface in the literature and at conferences it seems to me that at times we resemble scientists who have the means to study atoms of compound substances without being capable of proper separation and identification of the molecules they make up.

Here, perhaps, is my most basic disagreement with Kerlinger. He wants to study reading in the context of perception, motivation, attitudes, and so on. I believe that first of all I want to study perception, motivation, attitudes in the context of reading and, moreover, I don't want to study them in isolation from each other or from their organizing educational context.

Suppes (1974) in his analysis of the relationship between Foundational disciplines and education designates educational phenomena as the source of educational theory. However his treatment of the issue avoids the problems of fracturing of those phenomena as a function of the application of the conceptual and technical tools.

I am not certain that educational psychology, educational sociology etc. are capable of generating appropriate theories accounting for say, learning to read, in a public school context. They may need to develop a language and tools of their own beyond their current differentiation of their originating disciplines.

In classroom learning the interaction is the main event. Unless we make those interactions our subject of study we will have very little to say about reality and those who maintain that research has no contribution or only a very limited one to "the real world" are right.

In considering the relationship between basic research and the teaching and learning of reading it is important to remember that in the case of the physical sciences engineering intervenes between science and practical application.

About seven years ago Edmund Coleman in an elaborate proposal to the USOE sketched some ways in which the engineering concept might be applied to the translation of basic reading research and theoretical formulations into statements about the teaching of reading. In essence he proposed a stage in which variables which were discovered to have a relationship to the acquisition of reading skills be calibrated. As an example he constructed tables for the learnability (number of trials to criterion) of individual phoneme grampheme correspondences, for instance.

Another example of this engineering concept, is found in the development of readability formulae. The intolerable variations in readability of a particular passage between formulas, (I calculated the readability of 36 passages used by Coleman and Miller (1968) using five formulas (Flesch, Dale-Chall, Fry, Smog and Forecast). On some passages the range of estimates was as large as six years. One passage for example was rated grade 2 by one formula, grade 8 by another.) the variation between passages in a single text source (Bradley and Ames, 1977) and the problem of matching difficulty of text and ability of readers places the value of readability as currently conceptualized much in doubt.

Readability research is a prime example of applied research conducted without sufficient consideration for the educational context in which the application is to take place.

Coleman's proposal, however, should not be dismissed lightly. We have preciously little basic research and theory regarding either reading, the learning of reading or the teaching of reading. Attempts to directly translate from theory and/or basic research to practice are naive and lead to disappointments. Systematic applied engineering research may provide some answers in this respect.

Scriven, (1964) in an insightful essay notes that relative to such sciences as astronomy, psychology is in a very unfortunate position: it has to compete with common sense. And, he adds, fifty thousand years of common sense have resulted in a great deal of psychological knowledge being deposited in our language.

Educational research, and in particular reading research finds itself in a similar position. Everyone can read, everyone knows that reading can be difficult, that hard words make a story difficult, that long words are often hard words; that if there are many words in a story you can't pronounce chances are you won't understand it so well. Everyone knows, especially if they are CB'ers, that words only have meaning in defined communication contexts. And so forth.

So, what is left for psychology? As Scriven puts it:

> "Common sense not only steals the easy pickings from the field of study of human behavior, but it passes on to the science of psychology a set of extremely embarrassing questions because the everyday interactions of human beings leads them to need much more than their common sense knowledge quite frequently."

Scriven then goes on to say that though physicists are never asked to predict the fall of an individual leaf of a tree, psychologists are required to make predictions about the behavior of an individual learner at a point in time.

In my opinion Scriven, by this statement, questions the very existence of generalized theories of behavior, of the utility of a general psychology itself. One can view the possibility and utility of a generalized theory of a set of phenomena as a function of the degree of variation among the individual phenomena. It then becomes clear that a theory accounting for the reading process is less viable than a theory accounting for the behavior of gasses.

When faced with the demands of research in reading we often seem to retreat into the safety of translating common sense knowledge into pseudo psychological terms. A point in case is the swift adoption of the language of semantics and pragmatics currently in vogue for the same kinds of research questions expressed as little as five years ago in terms of Chomskian vocabulary. An analysis of the conference program of this year demonstrates this problem quite nicely. When this translation is a mere surface response, as in my judgement often is the case, I react as Archie Bunker to one of Meathead's learned speeches: "Why do you always have to use big words to hide that you don't know nothing?"

I view the proliferation of dependent variables upon which I commented a few years ago as a similar avoidance of reaction. (Editorial, *Journal of Reading Behavior,* Volume 6, Number 3)

Do we as reading researchers stay away from the classroom for fear of not being able to cope with the complexity of the phenomena?

One area where psychology, or at least psycholinguistics seems close to catching up with common sense is reading comprehension as a process. I believe that it is fair to say that the currently most widely held theoretical beliefs about reading comprehension are adequately summarized as follows:

Reading comprehension involves an interaction of some sort among decoding, language and knowledge of the world. The role of each of these factors depends on the level of the difficulty of the task involved, and the ability and orientation of the reader.

This is my brief summary of statements on the topic of Guthrie's recent book on comprehension, cognition and language (Guthrie, 1977).

I suggest that only a far more detailed analysis of the phrase "interaction of some sort" elevates this summary above the status of a commonsensical statement.

To offset the negative aspects of the proceeding comments I hasten to refer to my rather lengthy and optimistic treatment of the contributions I expect from lines of research exemplified by the work of Perry Thorndyke, (1976) among others (Tuinman, 1977). Now linguists have finally legitimized the study of semantics, pragmatics and rhetorical structure we should make some advances.

Where mere commonsense has not been able to guide us is in the development of effective techniques for teaching comprehension. Though psychometric research by Davis (1971) and others has empricially validated time honored pedagogical taxonomies there is very little teaching going on of main idea, of sequence, of predicting outcomes other than raw practice. After performing tasks related to these skills, discussion follows in terms of particulars, not in terms of the structure of the processes involved.

Perfetti (1977) perceives this when he remarks after summarizing his view of how people comprehend that of course, this doesn't tell us how people arrive at this ability nor how it should be taught. I paraphrase:

I call for basic research on the acquisition of comprehension skills by learners in an educational environment.

I see learning to read as a process which alternates between acquiring basic concepts about the structure of written language and application of these concepts in practice. The order and rate which these concepts are presented and learned depends in part on the teaching strategies employed.

At a very general level we know that learning to read becomes difficult for some children when they (a) fail to acquire these basic concepts and/or (b) fail to develop sufficient mastery in their application at any given stage.

For now I see a legitimate task for basic research into the learning-to-read process in a focus on the identification and analysis of such crucial concepts; an analysis of their relationship; of the development of automaticity in the application phase; of the limits of generalizability to specific pedagogical contexts. This kind of research has been neglected, I submit, for both the learning to decode and the development of comprehension, but more so for the latter.

I suggest that this task requires researchers sensitive to children's experiences in real classrooms. We do need a psychology of reading which is capable of explaining how people read with real books and how people learn to read in real classrooms.

If such a psychology is outdated when books and classrooms disappear or change in appearance, so be it. Psychology isn't timeless. It is an effort to understand reality as it is experienced now.

De Bono's (1976) second law states that scientific proof is usually no more than lack of imagination in providing alternative explanations. He refers to the unique explanations fitting one's data well as the Village Venus effect. The

villagers think their Venus is the most beautiful because their limited experience does not allow them to imagine a more beautiful girl.

If one accepts this law, and I cannot think of a reason why one shouldn't, the role of experience and creativity on the part of the researcher becomes clear. Truth becomes a matter of temporary lack of alternatives. Those lacking in their imagination are surest of their truths. I conclude in suggesting that immediate and intimate contact with readers is a prime source of alternative explanations for reading phenomena. I urge in particular those who are engaged in basic reading research to seek that contact.

REFERENCES

Arlin, M. One-study publishing typifies educational inquiry. *Educational Researcher,* 1977, *6*(9), 11-15.

Borgatta, E.F. Prologue: The current status of methodology in sociology. In E.F. Borgatta and G.W. Bohrnstedt, (Eds.), *Sociological methodology 1969.* San Francisco: Josey-Bass, 1969.

Bradley, J.M. and Ames, W.G. Readability parameters of basal readers, *Journal of Reading Behavior,* 1977 *9*(2), 174-184.

Bronfenbrenner, U. The experimental ecology of education. *Educational Researcher.* 1976, *5*(9), 5-15.

Buros, O.K. Fifty years in testing: Some reminiscences, criticisms and suggestions. *Educational Researcher,* 1977 *6*(7), 9-15.

Carroll, J.B. and Suppes, P. The committee on basic research in education: A four year tryout of basic science funding procedures. *Educational Researcher,* 1974 *3*(2). 7-10.

Comroe, J.H. and Dripps, R.D. Scientific basis for the support of biomedical science. *Science,* 1976, *192,* 105-111.

De Bono, E. *Practical thinking.* Baltimore: Pelican Books, 1976.

Davis, F.B. Psychometric research on comprehension in reading. *Reading Research Quarterly,* 1972, *8*(4), 628-678.

Fisk, M. What we've taught. In W.E. Brownson and J.E. Carter (Eds.), *Philosophical studies in education.* Proceedings of 1976 Annual Meeting of the Ohio Valley Philosophy of Education Society. Terre Haute: Indiana State University, 35-45.

Flesch, R. *The art of clear thinking.* New York: Harper and Row, 1951.

Gal'perin, P.J. Onderzoek van decognitieve ontwikkeling van het kind. *Peadagogische Studien,* Volume *49,* 1972.

Guthrie, J.T. (Ed.), *Cognition, curriculum and comprehension.* Newark, Delaware: IRA, 1977.

Hansen, C and Lowitt, T. *An applied behavior analysis approach to reading comprehension.* In J.T. Guthrie (Ed.), *Cognition, curriculum and comprehension,* Newark, Delaware: IRA, 1977.

Kerlinger, F.M. The influence of research on education practice. *Education Researcher,* 1977, *6*(8), 5-12.

Levy, B.A. Reading: Speech and meaning. *Journal of Verbal Learning and Verbal Behavior,* 1977, *16*(5), 623-638.

Meredith, G.P. The origins and aims of epistemics. *Instructional Science,* 1972, *1*(1), 9-30.

Perfetti, C. Language comprehension and fast decoding: Some psycholinguistic prerequisites for skilled reading comprehension. In J.T. Guthrie (Ed.), *Cognition, curriculum and comprehension.* Newark, Delaware: IRA, 1977, 20-41.

Petersen, P.E. *Die paedagogische Tatsachenforschung.* Besorgt von Theodor Rutt, Paderborn, 1965.

Scriven, M. Views of human nature. In T.W. Wann (Ed.), *Behaviorism and phenomenology — contrasting bases for psychology.* Chicago: University of Chicago Press, 1964, 163-190.

Suppes, P. The place of theory in educational research. *Educational Researcher,* 1974 *3*(6), 3-9.

Thorndyke, P.W. The role of inference in discourse comprehension. *Journal of Verbal Learning and Verbal Behavior,* 1976, *15*(4), 437-446.

Tuinman, J.J. *Reading comprehension: Recent developments.* Invited paper, Annual Meeting, International Reading Conference, Miami, Florida, May 1977.

Van Parreren, C.F. Het functioneren van leerresultaten. In C.F. van Parreren and K. Peeck. *Informatie over leren en onderwijzen.* Groningen (Neth.): Wolters, 1970.