DOCUMENT RESUME

ED 197 560 EC 131 767

AUTHOR Hansen, Cheryl L., Ed.: Haring, Norris G., Ed.
TITLE Child Assessment: The Process and the Product.
INSTITUTION Washington Univ., Seattle.
SPONS AGENCY Office of Special Education and Rehabilitative Services (ED), Washington, D.C.
PUB DATE Jul 80
CONTRACT 300-79-0062
NOTE 176p.: A Program Development Assistance System project.

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS Data Collection: Decision Making: Diagnostic Tests: *Disabilities: Elementary Secondary Education: *Evaluation Methods: Learning Disabilities: Recordkeeping: Referral: *Student Evaluation: *Testing: Test Selection

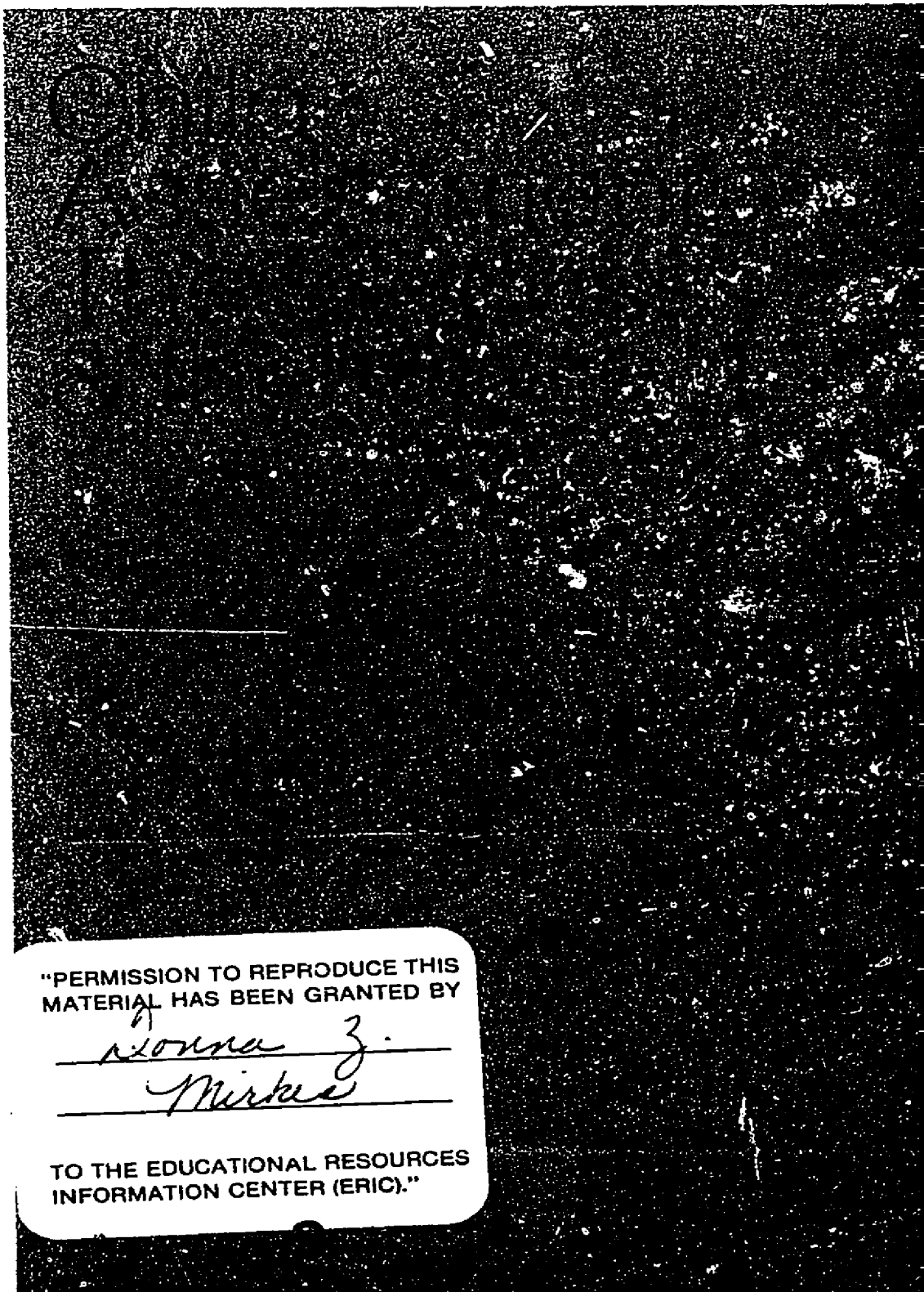
ABSTRACT

The document contains seven papers from the Child Assessment Topical Workshop designed to raise participant awareness of basic assessment issues involved in screening, placement, and measurement of daily performance: to provide specific information about the assessment process, including collecting, organizing, analyzing, and using data: and to provide an opportunity for model handicapped children's and special needs projects to review and refine their individual assessment processes and instruments. In "Current Assessment Practices: A New Use for the Susan B. Anthony Dollar?" B. Algozzine looks at problems and issues involved with assessment and data collection for learning disabled students. Decision making processes involved in student assessment and referral are the focus of "Basic Considerations in Child Assessment: Not Quite Everything You Wanted to Know...And More" by O. White. A third paper by C. McGuigan addresses "Selecting and Evaluating Educational Tests." "Assessing Social and Emotional Problems" is considered by B. Algozzine with particular emphasis on use of an observation scale. C. McGuigan, in a paper titled "Analysis and Use of Performance Data," discusses the establishment of aims, the identification of program guidelines and decision rules, and the analysis of data patterns to make appropriate educational or motivational interventions. C. Hansen points out in "Developing and Validating Assessment Instruments" that five variables should be taken into account in developing an assessment instrument: purpose of the assessment, what to assess, who will assess, how the assessment will be conducted, and whom the assessment results are for. The booklet closes with "Concluding Remarks: Using Assessment Data to Document Program Effectiveness" (C. Hansen). (SBH)

Editor
Cheryl L. Hansen

Norris G. Haring
Series Editor

PDAS
Program Development
Assistance System
University of Washington



Donna Z. Mirkes, Production Editor

Nicole M. Bussod, Editorial Assistant

Linda S. Floyd, Word Processor

Gregory M. Owen, Graphics Specialist

This document was produced under contract number 300-79-0062 from the United States Education Department, Office of Special Education and Rehabilitative Services, Division of Innovation and Development, Program Development Branch. The opinions expressed herein do not necessarily reflect the position or policy of the United States Education Department, and no official endorsement by U.S.E.D. should be inferred.

Printed in the United States of America.

July 1980

Contents

Preface

Assessment may be characterized as a process used to make educational decisions. It is more than a test and less than an entire program. However, assessment is the central cog around which programs revolve. Assessment is essential for programmatic decisions related to screening, placement, implementation and refinement. Accurate, reliable assessment information is critical to ensuring the appropriateness of educational programs for children. It is also critical to determining the worth of a particular program or instructional procedure. Finally, assessment is essential if education is to evolve from an art to a science.

Considering the importance of assessment to appropriate educational programming, it is frightening to reflect on its current status. People continue to make inappropriate decisions based on incorrect responses to inadequate questions. The questions are inadequate because the tests are too often invalid (Ysseldyke & Algozzine, 1979). The responses are incorrect because the tests are too often unreliable. The subsequent decisions might as well have been based on the flip of a coin.

One reaction to this problem is evident through the resurgence of informal teacher-made tests. A recent survey of federally funded model projects showed that 59% of the instruments used for assessment were teacher-made (Thurlow & Ysseldyke, 1979). Unfortunately, teachers are not trained to be test developers. While these instruments may serve immediate and practical needs for making educationally relevant decisions about children, they may not be acceptable as evidence of program effectiveness.

Due to the critical importance of identifying and using appropriate assessment instruments and in consideration of the problems associated with current assessment practices, the Office of Special Education identified assessment as the focus of a workshop for Handicapped Children's Model Programs and Special Needs Programs during 1980. These federally funded programs are mandated to develop, refine and replicate innovative, effective educational programs for handicapped children and youth throughout the nation. Thus, it is critically important that these programs apply the latest technology to develop assessment instruments. Further, these model programs must employ technically adequate instruments and procedures to demonstrate the effectiveness of their programs and to demonstrate that their programs improve the quality of education provided for children and youth.

Conducted by the Program Development Assistance System (PDAS), the Child Assessment Topical Workshop was held February 19-20, 1980 in San Antonio, Texas. PDAS is a federally funded technical assistance project mandated to assist model projects achieve their maximum potential. The participants at the workshop represented 20 model projects from around the nation.

Three major goals were identified for the Child Assessment Topical Workshop. These goals were 1) to raise participant awareness of basic assessment issues involved in screening, placement and measuring daily performance; 2) to provide specific information about the

assessment process, including collecting, organizing, analyzing and using data; and 3) to provide an opportunity for projects to review and refine their individual assessment processes and instruments.

Four methods were used to accomplish the workshop goals. First, large group sessions were held to present generic assessment information. These sessions became the basis of the present document. Second, assessment teams met to review individual assessment strategies and instruments and to share ideas for improving assessment plans. Third, each participant shared an assessment tool, battery or procedure currently used by his/her project. Finally, participants met individually with PDAS staff and the consultant to discuss unique assessment concerns.

A majority of the participants came to the workshop in search of a better assessment device. Some were dissatisfied with their current instruments and wanted to find new ones. Others did not know of an instrument which would provide needed information. Still others merely wánted confirmation that they had not overlooked a potentially useful tool. Those seeking easy answers to assessment concerns were quickly disillusioned. The perfect assessment instrument simply does not exist. Instead of pursuing the perfect instrument, participants were encouraged to concentrate on developing a range of assessment alternatives and to match them to specific assessment questions. Thus, participants were encouraged to approach assessment as a process rather than as a product.

REFERENCE LIST

Thurlow, M.L., & Ysseldyke, J.E. Current assessment and decision-making practices in model LD programs. Learning Disability Quarterly, 1979, 2, 15-24.

Ysseldyke, J.E., & Algozzine, B. Perspectives on assessment of learning disabled students. Learning Disability Quarterly, 1979, 2, 3-14.

Child Assessment

Current Assessment Practices: A New Use for the Susan B. Anthony Dollar?

Bob Algozzine[1]

Assessment of children takes many forms; broadly defined, it is the process of collecting data for use in making decisions about students. The activity of assessing may be differentiated from that of testing in that the latter may be defined as exposing a client to an instrument or set of questions primarily to obtain a score (Salvia & Ysseldyke, 1978), while assessment involves qualitative as well as quantitative data collection. Data obtained about a student may be used to make a variety of different educational decisions; Salvia and Ysseldyke (1978) differentiated five kinds of assessment related decisions. They indicated that assessment data are used in making decisions about screening, classification/ identification/ eligibility/ placement, instructional interventions, and pupil and program evaluations. Different types of information are necessary for different types of assessments. Collecting data about school students is an omnipresent activity; issues related to use of assessment data have been identified and discussed (Ysseldyke, 1978a; Ysseldyke & Algozzine, 1979).

Problems and concerns have evolved at each level of data collection and decision making with exceptional students.

When screening and classification decisions are made, definitional and conceptual issues are readily apparent, as are numerous issues regarding technical adequacy of devices and the extent to which assessment practices are biased. When data are used to plan instructional interventions, issues relative to the appropriateness of treatment modes as well as technical and practical adequacy are of interest. When data are used to evaluate pupil progress and program effectiveness, issues arise regarding both the nature of data to be collected and its relevance to different reference groups. That problems exist relative to current assessment practices is evidenced by a research project conducted to define psychometric characteristics which differentiate "classified" and "unclassified" underachievers.

The Twin Study[2]

Background. During 1978-79, the University of Minnesota's Institute for Research on Learning Disabilities (IRLD) conducted a study to determine whether typical assessment data would provide a basis for distinguishing between learning disabled (LD) students (as identified by the school district) and non-learning-disabled (non-LD) students who were experiencing academic difficulties (as indicated by a score below the 25th percentile in reading and/or math on the Iowa Tests of Basic Skills). If reliable differences could be found between the two groups on any of the assessment devices typically used in determining eligibility for services, these devices could be recommended for future use and the others discarded. If reliable differences could not be found, the time-consuming and costly use of these devices would be questioned and the need for better assessment devices or procedures recommended. The intent of the study was noble; the outcomes were provocative.

Procedures. Fifty LD and 49 non-LD students (from 30 individual schools in nine school districts in the greater Minneapolis/St. Paul metropolitan area) were administered assessment devices. Information on the student's cognitive ability, academic achievement, perceptual-motor skills and social affective skills (self-concept and behavior problems) was collected. The students were considered similar with regard to major demographic variables of interest; the demographic characteristics of the children in both groups are presented in Table 1. The assessment devices were those typically used in determining eligibility for LD services:

Bender Visual-Motor Gestalt Test
Developmental Test of Visual-Motor Integration (Beery)
Peabody Individual Achievement Test (PIAT)
Peterson-Quay Behavior Problem Checklist
Piers-Harris Self-Concept Scale

Stanford Achievement Test[3]
Weschler Intelligence Scale for Children - Revised (WISC-R)

Woodcock-Johnson Psycho-Educational Battery[3]

The scores of the LD and non-LD students on these devices were compared. In analyzing the data, raw scores were converted to standard scores when possible; otherwise, raw scores themselves were used for analysis.

Results. Statistical analyses of the test score data indicated that significant differences did exist between the LD and non-LD group means. Such differences were found on ten subtests of the Woodcock-Johnson Psycho-Educational Battery,[4] on the PIAT subtests, and on the Peterson-Quay Behavior Problem Checklist. On the Woodcock-Johnson and the PIAT, the LD group means on the subtests were below the non-LD group means. On the Peterson-Quay, where teachers rated students' problem behaviors, the LD group mean indicated a higher incidence of problem behaviors than did the non-LD group

TABLE 1

DESCRIPTION OF SUBJECTS FOR SELECTED DEMOGRAPHIC VARIABLES

			Age of Child (in months)		Parental Marital Status		Father's SES		Mother's SES		Family Income	
	Male	Female	Mean	S.D.	Married	Unmarried	Mean	S.D.	Mean	S.D.	Mean	S.D.
LD	40	10	121.04	5.04	26	9	58.32	25.84	47.56	24.16	$21423	10477
Non-LD	35	14	121.06	4.04	28	8	51.44	27.57	46.35	18.07	$22852	11027

mean. These statistically significant differences between group means generally reflected relatively small actual differences in mean scores. For example, for the Woodcock-Johnson subtests, the differences were from only 1.06 to 3.96 raw score points; for the PIAT subtests, mean score differences ranged from 4.94 to 8.89; for the Peterson-Quay, the mean score difference was 9.08.

The amount of overlap between the performance of LD students and that of non-LD students was derived by computing the percentages of scores from the two groups that were in a common range. The percentages of overlap between the LD and non-LD scores on the 49 individual measures ranged from 82 to 100%, with the median overlap being 96%. For example, on the PIAT mathematics subtest (on which a statistically significant difference between group means was found), overlap was 97%. This means that 97% of the scores obtained by LD students were within the same range of scores obtained by non-LD students.

Data were also analyzed by tallying the number of students in the two groups who earned identical scores. The number of identical scores (same score obtained by an LD student and a non-LD student) on the individual measures (excluding the Peterson-Quay Behavior Problem Checklist) ranged from 19 to 44; the number of identical scores possible was 49. For example, on the PIAT general information subtest (where the mean difference between groups was 7.75), 24 (49%) of the scores were obtained by both LD and non-LD students. On all but two measures, the number of identical scores was greater than 25 (51%). On the Peterson-Quay, data were not obtained on all students; with the number of identical scores possible being 33, 16 (48%) of the scores obtained by LD and non-LD students were identical. Results of statistical comparisons for selected psychoeducational devices are presented in Table 2.

Two additional analyses were performed to evaluate the effect of group similarities on standard classification

TABLE 2

STATISTICAL COMPARISONS FOR SELECTED PSYCHOEDUCATIONAL DEVICES

Domain	Test/Subtest	Non-LD Mean	Non-LD S.D.	LD Mean	LD S.D.	Mean Difference	Number of Identical Scores[b]	% Overlap
Cognitive	WISC-R Full Scale	102.88	9.72	99.92	12.66	2.96	27	99
	WISC-R Verbal	100.47	11.75	96.98	12.46	3.48	27	97
	WISC-R Performance	102.90	13.47	103.92	14.09	-1.02	22	98
	WISC-R Information[a]	101.94	11.63	96.30	11.42	5.64	39	99
	WISC-R Similarities	101.33	13.91	98.10	16.65	1.23	35	96
	WISC-R Arithmetic	95.10	10.97	93.10	10.44	2.00	42	100
	WISC-R Vocabulary[a]	102.55	11.14	97.20	10.40	5.35	33	93
	WISC-R Comprehension	106.22	12.35	102.86	15.24	3.36	35	98
	WISC-R Picture Completion	104.29	13.46	102.80	13.06	1.49	38	99
	WISC-R Picture Arrangement	106.63	12.72	106.90	16.34	-0.27	38	96
	WISC-R Block Design	98.78	17.57	102.50	13.33	-3.72	38	95
	WISC-R Object Assembly	105.51	14.62	107.55	17.65	-2.04	38	98
	WISC-R Coding	100.00	12.99	100.10	17.30	-0.10	37	98
Achievement	PIAT Math[a]	101.02	11.14	96.08	10.47	4.94	26	97
	PIAT Reading Comprehension[a]	100.51	7.34	93.04	11.01	7.47	31	92
	PIAT Reading Recognition[a]	100.69	8.42	91.80	8.98	8.89	19	90
	PIAT Spelling	95.84	8.17	88.48	10.33	7.36	25	92
	PIAT General Information[a]	104.31	9.10	96.56	10.38	7.75	24	90
	PIAT Total Test[a]	100.61	6.49	91.90	3.78	8.71	24	88
	Stanford Math Calculation	90.27	9.03	88.82	9.78	1.45	30	99
	Stanford Math Concepts	89.33	10.60	88.70	13.13	.63	31	99
Perceptual Motor	Bender	2.27	1.71	2.52	2.08	-0.44	44	99
	Beery	14.90	2.16	15.46	2.61	-0.56	39	99
Self-Concept	Piers-Harris	51.94	11.70	52.34	16.80	-0.40	21	97
Behavior Ratings	Behavior Problem Checklist[a]	10.21	10.40	19.29	15.22	-9.08	16	97

[a] Difference between means significant ($p < .05$).

[b] Number of identical scores possible was 49 except for BPC in which it was 33.

decisions. Three different indices of "severe discrepancy" were calculated for each child; in fact, differences between ability and achievement in five areas were obtained based on discrepancies greater than 1, 1.5, and 2 standard deviations. Classification by these three criteria was then compared to school classification. When a two standard deviation cut-off was used, only three of the 99 youngsters were "eligible" for LD status. When a one standard deviation cut-off was applied, 40 children were misclassified; and when a one and one-half standard deviation criterion was applied, a different 40 students were misclassified.

Discussion. Two groups of "categorically" different youngsters were studied; similarities were observed in the demographic as well as psychometric characteristics of the groups. In fact, many of the individuals within the groups obtained identical scores on the assessment devices. A direct comparison of PIAT mathematics performance scores for each group appears in Figure 1. The number of scores for which performance of LD and non-LD children was identical was 26 (i.e., 53%); to call these children "psychometric twins with different mothers" does not seem inappropriate or outrageous. That many of these children were misclassified is not surprising; who could tell them apart...?

Several interesting conclusions may be reached in analyzing the obtained results. Many professionals in the field of learning disabilities believe that current identification efforts miss many low achieving students who are, in fact, learning disabled, thereby resulting in denial of services to these students. That argument can be supported using the obtained data and subsequent analyses. One could very well argue that the students who were achieving poorly were, in fact, learning disabled. A case of reverse discrimination could be built for these youngsters; that is, no difference was observed in psychometric performance, yet only certain children received preferred treatment.

Others in the field might argue that a separate diagnostic category for "learning disabled" children is absurd when no apparent differences exist between school district underachievers and the special children. Statistically, of course, the "no difference hypothesis" is never proved; however, the nature of this study and the obtained results are suggestive as well as thought provoking.

Still other professionals in the field of learning disabilities would argue that too many students who are simply underachievers are identified as learning disabled and that such identification results in both stigma and limitation to students' life opportunities. This argument, too, can be supported by the obtained data. There were few psychometric differences in the performances of two groups of students. Using a 1.5 standard deviation deficit, 33 students were misclassified as LD, while only 7 were misclassified as non-LD.

It is little wonder that considerable confusion exists regarding identification of LD students. One needs only to pick his/her argument, then play with cut-off scores that will produce data to support the desired outcome. One conclusion from this study will generalize to all of special education; that is, there is considerable misclassification resulting from current assessment practices. In the IRLD study, 40% of the children were classified inappropriately when strict federal guidelines were used as performance criteria on which to evaluate school district decisions. That the psychometric assessments were expensive, time consuming and potentially harmful is evident; considerable expense, time and potential danger could have been spared by merely subjecting each of the 100 children to a coin flip (preferably of a Susan B. Anthony dollar for the most up-to-date evaluations) and using one or the other of the results (heads you're LD, tails you're not) take your chances. The probability of being LD is approximately the same as that obtained with a more "heavy-duty" evaluation.

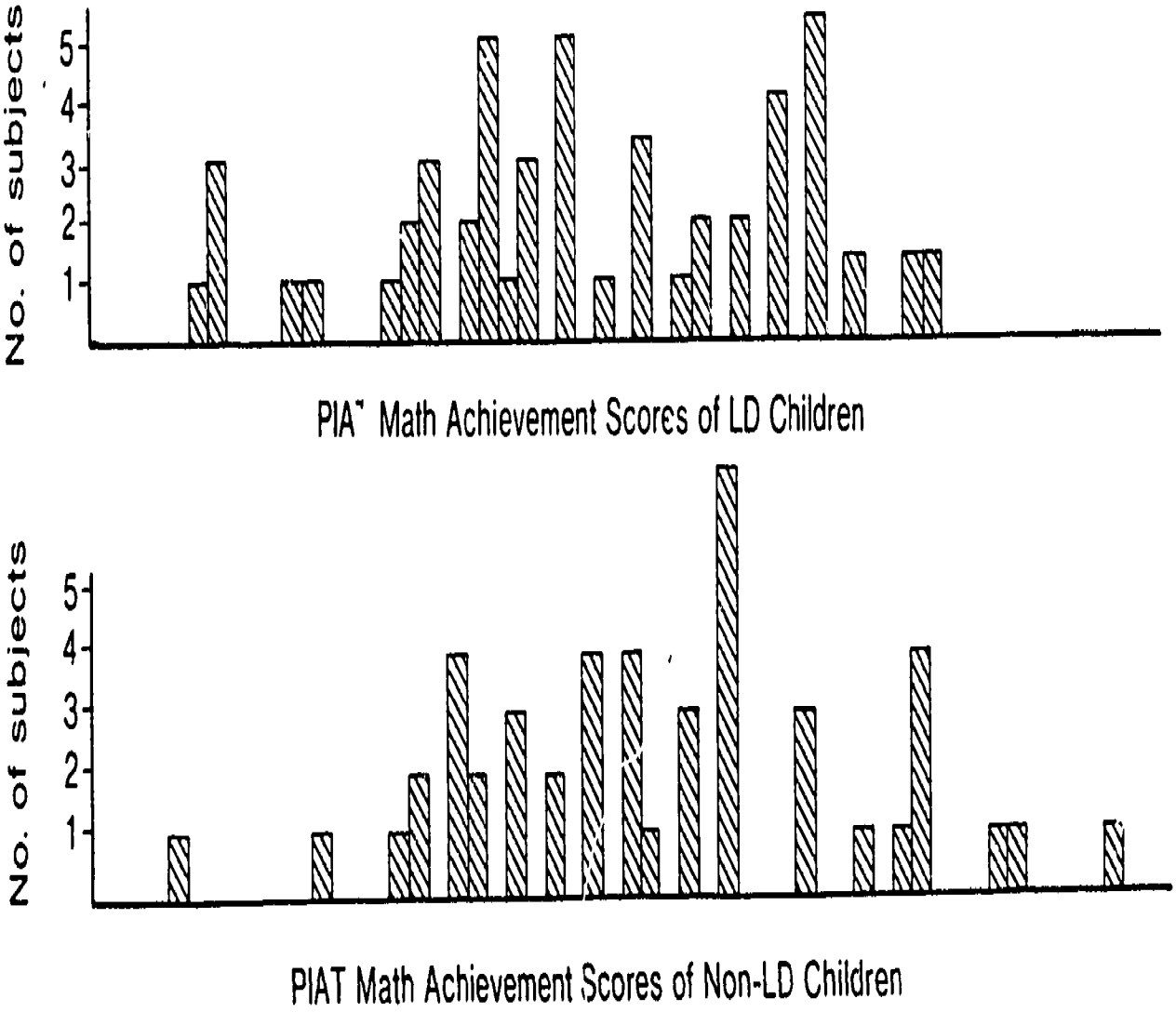


Figure 1: Distribution of Standard Scores on PIAT Math for LD and Non-LD Subjects

A discussion of other problems and related issues in the use of assessment data to make decisions about school-aged students seems warranted based on the outcomes of the "Twin Study". It is hoped that this discussion will stimulate users of assessment data to become critical, informed consumers of information about children. One thing is clear, current practices are anything but acceptable based upon these unresolved issues.

Critical Conceptual and Definitional Issues

Schools regularly collect considerable data on the students they serve. When students experience academic and/or social difficulties, school personnel regularly expand their data collection activities for those pupils. The data collected are intended to be useful in making psychoeducational decisions. The basis for much of the data collection is the belief that it will lead to identification of a special "disability" for which special treatment will become necessary. Unfortunately, few such "disabilities" exist.

For example, in spite of numerous attempts to create a more sophisticated condition, learning disabilities remains a category of underachievement. Algozzine and Sutherland (1977a) were critical of major components of then current definitions of learning disability. Specifically, they pointed out that psychological disorders articulated in definitions of learning disability were relatively obscure, that ability-achievement discrepancies were unreliable, and that little real evidence existed to support the notion of learning disability as a separate diagnostic category. The arguments made by these authors regarding the learning disabled category have been made by others (Quay, 1973; Reynolds & Balow, 1972) regarding other categorical labels.

Current practice in categorization of handicapped learners is often logically fallacious. A specific logical fallacy characterizes current identification efforts; the fallacy of an undistributed middle term, also called the fallacy of affirming the consequent, is the culprit.

The Fallacy of an Undistributed Middle Term. In its simplest form, the logical fallacy of an undistributed middle term follows a general paradigm in which a category or set of persons, places or things (A) covaries or coexists with another set of persons, places or things (B); a third category or set of persons, places, or things (C) is observed to coexist or covary with set B; set C is assumed identical to set A. Such reasoning is logical if, and only if, the relationship between sets A and B is both universal and specific; that is, when the characteristics in B appear in all (universal), and in only (specific) unit A. Obviously, this limitation restricts the utility of this form of reasoning. That the fallacy of an undistributed middle term permeates the field of special education can be readily illustrated.

Disorders or deficits said to be demonstrated by LD students are, for the most part, test named and test identified (i.e., auditory sequential memory deficits, figure ground pathology, grammatic closure disorders, body image problems, verbal expression disabilities and visual association deficiencies). Numerous statements appear in major textbooks and in the professional literature reporting that exceptional children (A) exhibit certain characteristics (B). The nature of the target characteristics controls the ease with which identification occurs; hence, alterations of intellectual criteria for mental retardation or levels of achievement discrepancy in learning disabilities result in alterations in prevalence. In fact, in special education we "get what we look for".

That we also engage in the practice of identifying children (C) who demonstrate the characteristics (B) listed in the textbooks should be apparent by our affinity for checklists, rating scales, "cut-off score" diagnoses and

profile analyses. Because of the nonspecific nature of those characteristics, we are often able to demonstrate that children do perform like the original individuals on whom the scale or test was "normed". Since the original relationship is not universal and specific, however, to a large extent our diagnoses (C is A) are often incorrect, unnecessary, and/or inappropriate. Special educators, then, engage in reasoning that follows a paradigm in which:

Disabled persons (A) exhibit certain behaviors (B).
Examinees (C) exhibit identical or similar behaviors (B).
Examinees (C) are disabled persons (A).

We might just as well engage in reasoning that concludes:

Cooters play in the mud.
Owen (an 8-year-old LD student) plays in the mud.
Owen is a cooter[5].

It should be clear that simply doing something that someone else does does not make you someone else; yet, the "undistributed middle term" in the assessment sequence has contributed to a tremendous lack of clarity and "false positive" identifications. Let the "twin study" be a reminder of the extent to which problems exist in identification practices.

The Fallacy of Affirming the Consequent. Put another way, identification practices within special education are subject to the "fallacy of affirming the consequent". In its simplest form, the argument follows this logic:

If the statement (A) is true, then a certain result will be observed (B).
Upon assessment, B is observed; it is then concluded that A is true.

If a child is disturbed, then the child will have certain characteristics.

Assessment results suggest the characteristics and it is concluded that the child is disturbed.

It is important to note that it is not the truth of the original statement that is at issue, but more the fact that the statement is not specific and universal. That is, it is not specified that the characteristics appear in only and in all disturbed children; there are clearly other reasons for the presence of the characteristics in question.

The ramifications of the definitional and conceptual issues underlying assessment of school-aged youngsters center on the problems of inappropriate or inefficient identification. Hallahan and Kauffman (1978) have suggested that the behavioral and other characteristics of mildly handicapped youngsters overlap (i.e., are not universal and specific) to a large extent. It is not difficult to hypothesize about reasons for that observation when one considers the basis for the identification in the first place (i.e., a logical non sequitur). Similarly, Hallahan and Kauffman (1978) suggested that categorically differentiated instruction is largely nonexistent. In spite of the truth of that contention, differential placements and classifications occur based upon the results of assessment efforts; this is largely due to the anticipated favorable outcome of treatment as opposed to the threat of unfavorable outcomes as a result of identification. One might also hypothesize that the reason professionals are willing to continue "affirming the consequent" is because occasionally the line of reasoning is correct. Therefore, like any other behavior put on an intermittent schedule of reinforcement, illogical reasoning is particularly resistant to extinction.

Bias Before, During, and After Assessment of LD Students. When making decisions about a group of students who are defined and described in as nebulous a manner as are most exceptional students, one must be concerned with subjective bias. In other words, whenever objective mechanisms for identification are absent, the probability is very high that subjectivity enters the decision-making process.

Much has been written about bias in the making of classification, identification, eligibility and placement decisions. Ysseldyke (1978a) summarized efforts in psychology to study bias in assessment, concluding that not only have psychologists been unable to agree on models or equations to be used in ascertaining test fairness, but they have been unable to agree on the concept of fairness. Educators are now repeating the mistakes made by researchers who have addressed bias, a fact readily apparent in concern for and attempts to identify the fair test for use with specific groups of students. In fact, bias occurs throughout the decision-making process, and is not restricted to bias in test usage.

Preassessment Bias. A variety of naturally occurring student characteristics have been shown to influence the formation of negative attitudes toward students. Facial appearance has been shown to influence placement decisions (Ross & Salvia, 1975), has been related to different personal and peer attitudes (Berscheid & Walster, 1974; Salvia, Sheare, & Algozzine, 1975), and has been shown to be a factor in differential teacher-pupil classroom interactions (Adams & Cohen, 1974; Algozzine, 1975). It has been demonstrated that other student characteristics (i.e., race, sex of child, achievement level of older siblings, socioeconomic status) differentially affect the formation and transmission of classroom teachers' expectations (Brophy & Good, 1974; Bergan & Smith. 1966; Carter, 1952; Coates, 1972; Datta, Schaefer, & Davis, 1968; Geisbrecht & Routh, 1979; Jackson & Lahaderne, 1967; Lenkowsky & Blackman, 1968; Lippett & Gold, 1959; Meyer & Thompson, 1956; Miller, McLaughlin, Haddon, & Chansky, 1968; Palardy, 1969; Rubovits & Maehr, 1973; Seaver, 1973).

It has also been demonstrated that behaviors of exceptional children result in differential teacher reactions. For example, Algozzine (1976, 1977) has shown that behaviors characteristic of emotionally handicapped (EH) youngsters were differentially bothersome to school personnel. Schlosser and Algozzine (1979) have shown

that behaviors characteristic of boys were more bothersome than those characteristic of girls, Mooney and Algozzine (1978) demonstrated that behaviors characteristic of LD children were less bothersome than those characteristic of EH children, and Giesbrecht and Routh (1979) found that the most influential category of information in teacher referrals was written comments concerning misbehavior. It seems, then, that even before a child utters one response to a test item, he/she may have the cards stacked unfavorably. The exact nature of this problem has not been specified. One possibility is that different assessment processes may be selected for different types of youngsters, or that examiners may hold preconceived notions about the outcomes of the assessment based upon the child's "characteristics".

Assessment Bias. If bias occurs before the assessment session, it may also occur during data collection and decision making. In fact, the circumstances of the testing, the influence of the examiner on the test results and observer biases have been studied in this regard; clearly, preassessment characteristics may be influential during the evaluation as well.

School psychologists often report that they receive referrals from certain teachers at a disproportionate rate when compared to others; similarly, they tend to base their decisions about the child on this fact (Hersch, 1971). The child's social class, appearance, parents' involvement in the school, referring teacher, reason for referral and other similar characteristics may result in differential interactions during testing sessions. Masling (1957) provides some evidence that the examinee's behavior during the evaluation may influence the outcome. In that study, two undergraduates were trained to respond in a warm, congenial manner, and a cool, aloof manner; examiners' interpretations of the Rorschach performances of the two subjects were more favorable when they behaved in an accepting (warm) manner. Neisworth, Kurtz, Jones, and Madle (1974) found that the diagnosis of hyperkinesis influenced observers' judgments about the child's behavior.

Expectations of an examiner, as cued by seemingly irrelevant characteristics of the testing circumstances and the examinee may be influential in assessment outcomes. These effects have been identified and studied in traditional assessment (i.e., individual interview evaluations) settings as well as in observational studies (cf. Hersen & Bellack, 1976; Stoneman & Gibson, 1978).

Postassessment Bias. In addition to bias that occurs prior to and during the collection of data for psychoeducational decisions, considerable bias occurs after the assessment as a function of the label assigned to the child. The labeling issue is relatively straightforward; what is of concern is what happens to the child as a result of the assignment of a categorical label. Labeling effects have been studied from two perspectives: 1) the impact of the label on the perceptions and behavior of the child, and 2) the impact of the label on others' perceptions and actions regarding that child (Algozzine & Mercer, in press). The general effects of labeling have been reported elsewhere (Goffman, 1963; Jones, 1977; MacMillan, Jones, & Aloia, 1974); a specific effect relates to the influence labels have on personal and interpersonal expectations for success and/or failure.

The special education labels have also been shown to be influential in biasing teachers' judgments (or interpersonal expectancies) about children. Interest in the effects of labeling probably stems from the work of Rosenthal and Jacobsen (1968) in which the experimenters attempted to generate differential student performances by biasing teachers. Within this context, the effects of manipulating various special education labels have been investigated in a variety of ways. Experimental studies have compared the effects of each disability label and have measured the labeling effects in teachers, undergraduate students and labeled youngsters. Selected investigations in which various labels have been studied are reported in Table 3. An analysis of the results of these investigations suggests that labels transmit negative expectations to teachers and other professionals likely to be working with handicapped

TABLE 3

SELECTED INVESTIGATIONS IN WHICH THE LD LABEL WAS STUDIED

Investigators	Label(s) Being Studied	Method of Investigation	Target Individual(s)	Results
Foster, Ysseldyke, 1976	LD vs ED vs MR vs N	Hypothetical and videotaped presentations--experimental comparisons	Transmission to teachers	More negative expectancies held for MR than for LD or ED; however, all special education categories viewed less favorably than normal one
Algozzine, Mercer & Countermine, 1977	LD vs ED	Hypothesized child was portrayed with label appropriate or inappropriate behaviors	Transmission to undergraduate teachers-in-training	Child thought to exhibit label-appropriate behavior was viewed differently than child thought to exhibit label inappropriate behavior
Algozzine & Sutherland, 1976 b	LD vs ED	Hypothetical child exhibiting aggressive behavior was rated in four case studies--experimental comparison	Transmission to undergraduate teachers in training	Child was viewed more favorably when thought to be learning disabled than when thought to be emotionally disturbed
Jacobs, 1978	LD vs N	Hypothetical and videotaped presentations--experimental comparisons	Transmission to classroom teachers	Labeled child was rated more negatively than non-labeled (i.e., normal) one
Mooney & Algozzine, 1978	LD vs ED	Characteristic behaviors of LD and ED children were rated	Transmission to vocational teachers	Behaviors of LD children were seen as less disturbing and bothersome than behaviors of ED children
Sutherland & Algozzine, 1979	LD vs N	Experimental study in which undergraduate students taught children labeled as LD or normal	Transmission to undergraduate student and to labeled or non-labeled child for production of effect	Performance of normal fourth grade children was differentially affected by label assigned to them prior to interaction with undergraduate "teacher"

children, and the effects of some labels are somewhat less negative (i.e., LD) than others (i.e., MR and ED).

It seems, then, that labels generate differential expectations and performances within interpersonal interactions as well as personal performances. Exceptional children have also been shown to be rejected by their peers and to be recipients of less desirable teacher and peer interactions (Bryan, 1974, 1977; Bryan & Bryan, 1978). The research on bias following assessment suggests that a special education label affects the lives of children who receive it. The effects suggest that this influence goes beyond simply making the child eligible for (and/or providing) special educational treatment.

Technical Adequacy

One of the most critical issues in making psychoeducational decisions for and about LD students is that the standardized tests used are often technically inadequate. Ysseldyke and Salvia (1974) provided a list of reliabilities for commonly used norm-referenced tests, reporting that the majority were technically inadequate. Salvia and Ysseldyke (1978) listed tests with inadequate reliability and validity. Ysseldyke, Algozzine, Regan, & Potter (1979) demonstrated that decision makers use inadequate tests as often as they use adequate ones. To some extent, then, decision making is characterized by the use of information derived from tests with less than adequate technical characteristics.

To compensate for the fact that most tests lack perfect reliability, the standard error of measurement may be a useful addition to decision making. Salvia and Ysseldyke (1978) suggest that estimating true scores and building confidence intervals around them may be an appropriate method of reducing some of the uncertainty inevitable in the use of imprecise measures. The extent to which that

recommendation will be heeded is a question for future research; historically, decision making for identification/ placement and instructional programming has been based on the assumption that the obtained score is the true score. Few practitioners acknowledge that the sample of behaviors represented by the score is limited and that subsequent assessment may result in a different score.

Appropriateness of Treatment Models

The debate over what should be tested and taught within the assessment-intervention paradigm has been around for quite some time (cf. Salvia & Ysseldyke, 1978). It boils down to a decision about what should be evaluated as a basis for effective programming. Ysseldyke and Salvia (1974) suggested that two competing viewpoints have dominated this controversy; ability training and task analysis are the names assigned to each position.

Those who advocate the ability training point of view believe that there are specific abilities which underlie the acquisition of academic skills and that for most children failure to acquire these skills is a direct result of ability deficits. The argument goes something like this:

Disabled learners have perceptual problems.
The target child has perceptual problems.
The target child is a disabled learner.

It is then reasoned that correction of the perceptual problem will result in correction of the learning disability. Some support for this model has been compiled; however, success is clearly doomed to logical bounds.

Those who advocate a task analysis point of view contend that specific abilities do not underlie academic success, but more, specific subskills underlie academic success. Proponents of this model contend that failure at academic

tasks results from failure to learn necessary prerequisite skills. The argument goes something like this:

Students with reading difficulties have skill deficits in reading.
The target child has a skill deficit in reading.
The target child is a student with reading difficulties.

It is then reasoned that correction of the skill deficits will result in correction of the reading problem. Some support for this model has been compiled. Success is again bounded by the logical nature of the argument; that is, the argument is logical if, and only if, the skill deficit is both universal and specific.

In addition to the task analysis vs. ability training controversy, issues related to the practical use of assessment information have arisen. Many assessment instruments are useful for making some decisions, but are quite limited relative to others. The Peabody Individual Achievement Test (PIAT), for example, may be helpful for screening or program placement (identification) decisions. The format of score representation (e.g., grade equivalents, percentiles) provides global measures of academic functioning. The PIAT is inadequate with regard to educational planning; it is difficult to program for a child on the basis of age or grade equivalency scores (i.e., what do you teach a child who earns a 3.2 in mathematics?).

Teachers and diagnosticians administer a wide variety of tests in which only developmental scores (e.g., age or grade equivalencies, percentiles) result. Analysis of test performance on these measures tells them little about the child's knowledge relative to the items sampled. A method for alleviating this problem has begun to emerge; in fact, diagnostic testing involves identifying specific strengths and weaknesses in performance as well as obtaining a global performance score. Knowledge of the problems which a child solves and does not solve and their

content differentiates the diagnostic approach to testing. Any test can be made diagnostic; it is simply necessary to identify the content (or behaviors sampled) of each item and then develop a form to tabulate correct and incorrect answers. Analysis of errors within test performance can then become the basis for further assessment and educational programming. Algozzine and McGraw (1980) have applied this model to the PIAT Mathematics Subtest. It should be obvious that diagnostic testing and error analysis provide more information for educational programming than simply recording global performance scores.

Regardless of the position taken relative to current "best practices" in utilizing assessment information to build instructional programs, problems are apparent. Clearly, the more important consequence of assessment is treatment; however, with the problems evident in assessment, one wonders why it is not possible simply to provide more effective treatment without the unnecessary inconvenience of a diagnostic evaluation.

Nature of the Data Collected

When evaluating pupil progress in a treatment program, or when evaluating that program's effectiveness, the weaknesses of global scores of improvement and statistical analyses become apparent. For the most part, performance on norm-referenced and many criterion-referenced measures is based on the number of raw score points earned, i.e., the number of items correct. The traditional pretest/posttest evaluation model may be sensitive to shifts in global scores derived from items correct and incorrect, but special conditions are often necessary for achievement of "significant gains". Statistical significance is a function of group variability and size as well as magnitude of difference in performance. When a large number of subjects is

evaluated and when those subjects are similar in performance, a relatively small difference between the groups may be "significant" (the "twins" reappear). Similarly, however, when n is small and group variability is large, relatively large differences may not be statistically significant. To combat and/or not misuse this phenomenon, program evaluators have begun to set a priori criteria on which to base effectiveness; statisticians have always recommended "power analyses" to aid in decision making relative to this problem area. In practice, this means that it is appropriate to evaluate program progress against pre-set criteria in spite of (or in addition to) the statistical significance of other results; selection of the criteria should be based on meaningful progress estimates. To accomplish this, an evaluator should collect data on the historical (or baseline) progress of the target indivdiual or group and/or some control unit; improvement can then be judged against the established performance record. The "n of 1" research designs offer excellent alternatives on which to base such comparisons.

Relative to program evaluation and pupil progress monitoring, the sensitivity of the measurement also becomes critical. When setting criteria for improvement in a program evaluation or a child progress evaluation, the following objective may be delimited: "To gain a year in reading after a year in the model program instruction". While the objective may be appropriate in terms of improved performance, traditional measurement practices make it difficult to attain; psychometric devices are often not sensitive to use at this level. To make a year's progress, a certain number of items must be answered correctly in addition to the number previously answered correctly; improvement, then, is a function of the items correct. The items in the test may or may not be sensitive indicators of the model program content.

A Plea

This material was not written to be an instructional module in improved assessment practices; it was intended to expose some issues of concern relative to screening, identification, intervention and progress and/or program evaluation. It should be evident that problems exist in current assessment practices. The controversies over bias, test selection, item use, and statistical significance or importance will rage on long after the last psychometrician has tossed in his/her Binet and PIAT. Of major concern to educators should be the extent to which assessment results are turned into practice; that challenge remains....

FOOTNOTES

[1]Sections of this material appear in articles or reports coauthored by Bob Algozzine and James Ysseldyke; contact Bob Algozzine, G325C Norman Hall, University of Florida, Gainsville, FL 32611 for additional information or reprints.

[2]This research was performed pursuant to a contract from the Office of Special Education, United States Education Department to the Institute for Research on Learning Disabilities, Department of Psychoeducational Studies, University of Minnesota, Contract #300-77-0491. Bob Algozzine is affiliated with the Institute; portions of the twin study description were taken from a summary prepared by James Ysseldyke, Director of the Institute. Complete research reports are available from the IRLD Editor, 350 Elliot Hall, University of Minnesota, Minneapolis 55455 (RR #13).

[3]Selected subtests of these devices were administered.

[4]Of the ten subtests showing statistically significant differences, two were from Part One: Tests of Cognitive Ability (Memory for Sentences, Antonyms-Synonyms), and eight were from Part Two: Tests of Achievement (Letter-Word Identification, Word Attack, Passage Comprehension, Dictation, Proofing, Picture Vocabulary, Quantitative Concepts, Applied Problems).

[5]Mere usage of terms to connote a condition should not be interpreted as endorsement of those terms as new labels for the condition.

REFERENCE LIST

Adams, G., & Cohen, A. Children's physical and interpersonal characteristics that effect student-teacher interactions. The Journal of Experimental Education, 1974, 43, 1-5.

Algozzine, R. F. Attractiveness as a biasing factor in teacher-pupil interactions. Unpublished doctoral dissertation, The Pennsylvania State University, 1975.

Algozzine, B. The disturbing child: What you see is what you get? The Alberta Journal of Educational Research, 1976, 22, 330-333.

Algozzine, B. The emotionally disturbed child: Disturbed or disturbing? Journal of Abnormal Child Psychology, 1977, 5, 205-211.

Algozzine, B. & McGraw, K. Diagnostic testing in mathematics: An extension of the PIAT? Teaching Exceptional Children, 1980, 12, 71-77.

Algozzine, B., & Mercer, C. D. Labels and expectancies for handicapped children and youth. In D. A. Sabatino & L. Mann (Eds.), Fourth review of special education. New York: Grune & Stratton, in press.

Algozzine, B., Mercer C. D., & Countermine, T. The effects of labels and behavior on teacher expectations. Exceptional Children, 1977, 44, 131-32.

Algozzine, R. F., & Sutherland, J. Non-psychoeducational foundations of learning disabilities. The Journal of Special Education, 1977, 11, 91-98. (a)

Algozzine, B., & Sutherland, J. The "learning disabilities" label: An experimental analysis. Contemporary Educational Psychology, 1977, 2, 292-297. (b)

Bergan, J. R., & Smith, J. O. Effects of socioeconomic status and sex on prospective teacher judgments. Mental Retardation, 1966, 4, 13-15.

Bersheid, E., & Walster, E. Physical attractiveness. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 7). New York: Academic Press, 1974.

Brophy, J. E., & Good, T. L. Teacher-student relationships: Causes and consequences. New York: Holt, Rinehart and Winston, 1974.

Bryan, T. Peer popularity of learning disabled children. Journal of Learning Disabilities, 1974, 7, 621-625.

Bryan, T. Learning disabled children's comprehension of nonverbal communication. Journal of Learning Disabilities, 1977, 10, 501-506.

Bryan, T., & Bryan, J. Understanding learning disabilities (2nd ed.). Port Washington, N.Y.: Alfred, 1978.

Carter, R. How invalid are marks assigned by teachers? Journal of Educational Psychology, 1952, 43, 218-228.

Coates, B. White adult behavior toward black and white children. Child Development, 1972, 43, 143-154.

Datta, L., Schaefer, E., & Davis, M. Sex and scholastic aptitude as variables in teachers' ratings of the adjustment and classroom behavior of Negro and other seventh grade students. Journal of Educational Psychology, 1968, 59, 94-101.

Foster, G. G., & Ysseldyke, J. Expectancy and halo effects as a result of artifically induced bias. Contemporary Educational Psychology, 1976, 1, 37-45.

Giesbrecht, M. L., & Routh, D. K. The influence of categories of cumulative folder information on teacher referrals of low-achieving children for special educational services. American Educational Research Journal, 1979, 16, 181-187.

Goffman, E. Stigma: Notes on the management of a spoiled identity. Englewood Cliffs, N.J.: Prentice Hall, 1963.

Hallahan, D., & Kauffman, J. Labels, categories, behaviors: ED, LD, and EMR reconsidered. Journal of Special Education, 1978, 11, 139-147.

Hersen, M., & Bellack, A. S. Behavioral assessment: A practical handbook. Elmsford, N.Y.: Pergamon Press, 1976.

Hersch, J. B. Effects of referral information on testers. Journal of Consulting and Clinical Psychology, 1971, 37, 116-122.

Jackson, P., & Lahaderne, H. Inequalities of teacher-pupil contacts. Psychology in the Schools, 1967, 4, 204-211.

Jacobs, W. R. The effect of the learning disability label on classroom teachers' ability objectively to observe and interpret child behaviors. Learning Disability Quarterly, 1978, 1, 5-55.

Jones, R. A. Self-fulfilling prophecies. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Lenkowsky, R., & Blackman, L. The effect of teachers' knowledge of race and social class on their judgments of children's academic competence and social acceptability. Mental Retardation, 1968, 6, 15-17.

Lippett, R., & Gold, M. Classroom social structure as a mental health problem. Journal of Social Issues, 1959, 15, 40-49.

MacMillan, D. L., Jones, R. L., & Aloia, G. F. The mentally retarded label: A theoretical analysis and review of research. American Journal of Mental Deficiency, 1974, 79, 241-261.

Masling, J. The effects of warm and cold interaction on the interpretation of a projective protocol. Journal of Projective Techniques, 1957, 21, 377-383.

Meyer, W., & Thompson, G. Sex differences in the distribution of teacher approval and disapproval among sixth grade children. Journal of Educational Psychology, 1956, 47, 385-396.

Miller, C., McLaughlin, J., & Chansky, N. Socio-economic class and teacher bias. Psychological Reports, 1968, 23, 806.

Mooney, C., & Algozzine, B. A comparison of the disturbingness of LD and ED behaviors. Journal of Abnormal Child Psychology, 1978, 6, 401-406.

Neisworth, J. T., Kurtz, P. D., Jones, R. T., & Madle, R. A. Biasing of hyperkinetic behavior ratings by diagnostic reports. Journal of Abnormal Child Psychology, 1974, 2 (4), 323-329.

Palardy, J. What teachers believe - what children achieve. Elementary School Journal, 1969, 69, 370-374.

Quay, H. C. Special Education: Assumptions, techniques, and evaluative criteria. Exceptional Children, 1973, 49, 163-170.

Reynolds, M. C., & Balow, B. Categories and variables in special education. Exceptional Children, 1972, 38, 357-366.

Rosenthal, R., & Jacobsen, L. Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. New York: Holt, Rinehart, & Winston, 1968.

Ross, M. B., & Salvia, J. Attractiveness as a biasing factor in teacher judgments. American Journal of Mental Deficiency, 1975, 80, 96-98.

Rubovits, P., & Maehr, M. Pygmalion black and white. Journal of Personality and Social Psychology, 1973, 25, 210-218.

Salvia, J., Sheare, J., & Algozzine, B. Facial attractiveness and personal social development. Journal of Abnormal Child Psychology, 1975, 3, 171-178.

Salvia, J., & Ysseldyke, J.E. Assessment in special and remedial education. Boston: Houghton Mifflin, 1978.

Schlosser, L., & Algozzine, B. The disturbing child: He or she? The Alberta Journal of Educational Research, 1979, 25, 30-36.

Seaver, W. B. Effects of naturally induced teacher expectancies. Journal of Personality and Social Psychology, 1973, 28, 333-342.

Stoneman, Z., & Gibson, S. Situational influences on assessment performance. Exceptional Children, 1978, 45, 166-169.

Sutherland, J. H., & Algozzine, B. The learning disabilities label as a biasing factor in the visual-motor performance of normal children. Journal of Learning Disabilities, 1979, 12, 17-23.

Thurlow, M. L., & Ysseldyke, J. E. Current assessment and decision-making practices in model programs for learning disabled students. Learning Disabilities Quarterly, 1979, 2, 15-24.

Ysseldyke, J. E. Diagnostic-Prescriptive teaching: The search for aptitude-treatment interactions. In L. Mann & D. Sabatino (Eds.), The first review of special education. New York: Grune & Stratton, 1973.

Ysseldyke, J. E. Implementation of the nondiscriminatory assessment provisions of Public Law 94-142. In Developing criteria for the evaluation of protection in evaluation procedures provisions. Washington, D.C.: Department of Health, Education, and Welfare, United States Office of Education, Bureau of Education for the Handicapped, 1978. (a)

Ysseldyke, J. E. Remediation of ability deficits in learning disabled adolescents: Some major questions. In L. Mann, L. Goodman, & J. L. Wiederholt (Eds.), The learning disabled adolescent. Boston: Houghton-Mifflin, 1978. (b)

Ysseldyke, J. E. Issues in psychoeducational assessment. In D. Reschly & G. Phye (Eds.), School psychology: Methods and roles. New York: Academic Press, 1979.

Ysseldyke, J. E., & Algozzine, B. Perspectives on the assessment of learning disabled students. Learning Disabilities Quarterly, 1979, 2, 3-13.

Ysseldyke, J. E., Algozzine, B., Regan, R., & Potter, M. Technical adequacy of tests used in simulated decision making (Research Report No. 9). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities, 1979.

Ysseldyke, J. E., & Salvia, J. A. Diagnostic-prescriptive teaching: Two models. Exceptional Children, 1974, 41, 181-186.

Basic Considerations in Child Assessment: Not Quite Everything You Wanted to Know . . . and More[1]

Owen R. White

"Child assessment" has been used by educators to mean a variety of things ranging from any attempt to measure or quantify something about children to full-blown program evaluations. It seems particularly common for people to confuse "assessment" with initial "data collection" and "child assessment" with "program evaluation" (Anderson, Ball & Murphy, 1975). It might be helpful, therefore, to begin with a brief overview of those terms before taking a closer look at child assessment per se.

Assessment or evaluation in any form is considerably more than simple data collection. After the data are gathered, they must be organized and treated in a way which will actually help someone make a prespecified decision. If the data are ignored, or do not prove useful in making the particular decision desired, then evaluation or assessment in the true sense of the word really has not occured.

The difference between child assessment and program evaluation lies in the type of decision one wishes to make[2]. In program evaluation the decisions involve the

development, implementation, refinement and/or termination of general programs designed to serve groups of children. For example, program evaluations might be set up to help decide what buildings and facilities need to be constructed, what staff must be hired, how the staff should be trained, or what general curriculum materials should be purchased. In some cases the decision will affect all children in a program (e.g., all children might be placed in the same facility), while in other cases decisions will affect only subgroups of children (e.g., not all children will require all of the special service options a program offers). In each case, however, the overall purpose of program evaluation is to help the educator make informed choices while deciding how to set up a general framework for meeting the needs of some target population.

In child assessment the focus shifts from the overall effectiveness of a general program to the specific needs of each individual child. Perhaps, for example, a certain approach to teaching self-help skills works with most children, but Billy needs something different. Most children in a particular program might progress well with only a half-hour of special help, but the teacher decides to give Susie more time. The general curriculum for reading begins with letter sounds, but Jose might be able to start at a much higher level in the sequence. Within the general framework established through careful program evaluation, the purpose of child assessment is to help in deciding which specific options should be employed for each child and, when none of the general options appear adequate, to guide in the development of new programs which will meet individual needs.

Some of the elements of a system designed to facilitate an overall program evaluation might be borrowed from an existing child assessment system or vice versa. Perhaps a test used to evaluate overall program success is also useful for deciding which pupils should be transferred to another program. Overlap between systems should be encouraged. For at least two reasons, however, it is

unlikely that a child assessment system would adequately meet all program evaluation needs or vice versa.

First, in most program evaluations, it is desirable to collect comparable data on all children within the program so a comprehensive and meaningful picture of overall program success can be formulated. When making decisions concerning individual pupils, on the other hand, the system must be flexible enough to allow the collection of whatever information is of most use in that individual case -- even if some or all of that information would be of little or no use in making decisions for other children.

Second, appropriate child assessment will frequently lead to rapid program changes and modifications, whereas meaningful program evaluation usually requires that the program be consistently applied without change over some predetermined period. If a pupil appears to be having difficulties during the first few days of a new vocational program, child assessment might prompt a change in the program, but if the focus were on the evaluation of the program itself, the program might be continued without change in order to give it every opportunity to work. In short, even if the same basic information is being collected, the functional outcomes of child assessment and program evaluation can still be in virtual opposition to one another. The role each system plays in the development and conduct of a program is entirely different from that of the other. This chapter will focus only on child assessment.

THE ROLE OF CHILD ASSESSMENT IN MODEL PROGRAMS

Most educators limit their concern for formal, well documented child assessment systems to areas in which consistently appropriate decisions are difficult to reach. If, for example, the staff in a particular program seem to have some difficulty in deciding exactly where a pupil's instruction in a curricular sequence should begin, some attempt might be made to develop a more precise approach for making those decisions. If an informal approach appears to be working in some other area of the program, there is a tendency to leave it alone. After all, trying to impose additional structure in an already successful system might do more harm than good, and even if a new system would not disrupt the established pattern of success, the resources used to develop that system might be spent more profitably elsewhere. Still, when adopting such an attitude, it is important to realize that the decision to leave things alone is almost certainly going to be influenced by individual skills and competencies. Other people might not enjoy the same success as current staff and previously unnoticed problems may come to light when someone leaves or a new person is hired. That may be of only limited concern with some programs, but it should be of paramount importance to the type of model program for whom this book is written.

Model programs are not funded simply to serve the pupils in a given locality. They are funded to demonstrate an exemplary approach to the education of some target population and to work actively for the adoption and replication of that approach by other programs working with similar children. Even if the model program's staff does not appear to have any problems with a particular type of decision, a careful analysis of the competencies and skills which might be responsible for that success

should still be completed. If the analysis reveals the need for skills and competencies unlikely to be shared by the staff of potential adopter programs, then the development of formal assessment systems to offset those deficiencies should be seriously considered. The role of child assessment systems in model programs is not limited to making timely and appropriate decisions concerning individual pupils. It extends to making the basic program itself easily and efficiently replicable.

In order to develop child assessment systems that are sufficiently precise to enable replication, many factors need to be considered. Each set of factors is discussed in some detail below.

WHEN IS A DECISION A DECISION?

Since the purpose of child assessment is to help the educator make decisions about individual pupil programs, the development of a child assessment system must begin with the specification of the decisions which need to be made concerning each child. Some care is required, for if this step in the process of system development is not properly completed, the whole system will be directionless or, perhaps worse, misdirected. For example, one might state that we want to decide which of a pupil's performance deficits are greatest. That seems reasonable, but knowing a pupil's performance deficits will not necessarily result in any particular action concerning his or her program. Instruction might begin with skills which relate to those needs, but work might just as easily proceed with other, seemingly less important skills: prerequisite skills, skills of immediate concern to the parents, or even skills which are in themselves unimportant, but can serve as simple instructional targets to help the teacher and pupil "get to know one another."

"Deciding" which of a pupil's needs are greatest amounts to little more than arriving at some conclusion, discovering some fact or bit of information, or making some statement of belief. Many of the presumed purposes of child assessment fall into that category, but while that information might have tremendous impact on the pupil's program, it might not. It is important, therefore, to make a distinction between "determinations" (the process of documenting or discovering something) and "true decisions" (the process of making a choice among possible actions which will directly affect the future course of a child's program). One may lead to the other, but in the long run it is the decision -- the action -- that counts. A few examples might illustrate the point more clearly:

> One might try to determine a child's skill development needs as one piece of information useful in deciding where instruction should begin.
>
> If one could determine whether a child had a hearing deficit, it might be easier to decide which of several communication programs to try first.
>
> By determining whether a pupil is having trouble in a regular class, the process necessary for deciding if the pupil should be in a special education program can begin.

In other words, "determinations" may be a necessary part of the process leading to a decision, but they rarely represent the decision in and of themselves. A decision, in the truest sense of the word, must involve some explicit choice among possible actions. Unless the actions which might be taken are described, the decision itself has not been clearly defined. Unless the decision to be reached is clear, meaningful assessment cannot occur.

WHEN IS A DECISION APROPRIATE?

After ensuring that an assessment system leads to some clearly defined decision (choice among actions), the next step in developing or validating the system is to determine whether the decisions reached are consistently appropriate. After all, the purpose of child assessment is to help the pupil, not simply to make changes.

The question of whether a decision is "appropriate" is complex and subject to a number of qualifications. It is simple enough to say that a programmatic decision is appropriate to the extent that it results in an improvement in the pupil's life (Stufflebeam, Foley, Gephart, Guba, Hammond, Merriman & Provus, 1971), but defining "improvement" and weighing the relative merits of different sorts of improvement may be somewhat difficult. The Education for All Handicapped Children Act (PL 94-142) distributes the responsibility for making those determinations among several people, including teachers, local administrators, relevant specialists, parents and, when possible, the pupil. Generally, it is assumed that if all concerned can agree that a decision was important and appropriate, it was. Still, the dynamics of reaching a group decision are frequently less than satisfactory and often subject to inappropriate influences such as personal pressure, availability of programs and resources, the pupil's sex or race, and the vested interests of social agencies and advocacy groups (Holland, 1980). Clearly, the resolution of those problems is beyond the scope of this chapter. It may be possible, however, to outline at least a few considerations involved in determining whether an assessment process has led to an appropriate decision.

Borrowing from traditional testing theory, one might assume that a decision is valid if it accomplishes the intended purposes(s) (English & English, 1958; Anderson Ball & Murphy 1975). Before an assessment process is

implemented, therefore, it would seem wise to obtain a clear consensus of opinion concerning the desired outcome and then to test the validity of the process through an examination of the actual outcome. For example, the list of annual goals on a pupil's IEP are supposed to reflect the skills which the child study team feels the pupil can and should develop during the coming year. If the pupil eventually meets those goals, some people might be willing to assume that the child assessment system which led to their selection is working. Perhaps, however, the assessment system only led to the selection of goals which would be easy to reach. Are there other areas of development in which the pupil failed to make progress because no goals were set? Did reaching the specified goals result in some improvement in the pupil's status (like helping the pupil move to a less restrictive environment)? Obviously, there are a number of different criteria by which the selection of goals might be judged, and unless the relative importance of those criteria are discussed ahead of time, it will be difficult or impossible to assess the degree to which the assessment system led to "appropriate" decisions.

After arriving at some consensus of opinion concerning the criteria for an appropriate decision, educators are frequently dismayed to learn that decisions did, indeed, fail to accomplish the desired results (e.g., the pupil failed to reach the intended goals). Before questioning the validity of the decision itself, however, one should first determine whether the result of the decision was actually carried out. Each decision should result in some action. Did it? If, for example, an IEP team decided that a program must be developed in a particular curricular area, it should be reasonable to assume that the program was, in fact, developed and implemented. According to Alper (Note 1), that may not be a safe assumption at all. In a study of the IEP process in California, Alper and his associates discovered that more than 50% of the programs described on individual pupil IEPs were not implemented at any observable level in the classroom. If the classroom teachers found it necessary to ignore the IEP goals

because they were based on faulty information or goal selection procedures, then the assessment process should be changed. If, on the other hand, the failure to implement programs for certain goals was due to a lack of skill or motivation on the part of the teachers, then teacher training or other action should be considered -- the assessment process itself might be left alone, at least until other evidence is gathered that suggests the need for a change.

The reliability of the decision-making process can also influence the validity of a decision. What Salvia and Ysseldyke (1978) say about tests should also be true of decisions: "Reliability is a necessary but not a sufficient condition for valid measurement...the reliability of a test limits its potential validity (p.104)." As applied to assessment decisions, that might be translated to mean that if the same basic assessment information, treated in the same basic way, does not lead consistently to the same decision, then the decision itself may be invalid. The word may needs to be emphasized, however. The relationship of a test to a test score is considerably more direct than the relationship between a body of information and an educational decision. While one might assume that the same test, given to the same pupil, should always yield the same information regardless of the person administering the test, we cannot divorce the role of that person in the total decision-making process. Different people will have different backgrounds and experiences with the pupil and may, quite legitimately, arrive at different conclusions.

Still, to the degree that an assessment process does actually guide decisions in a clear-cut, replicable fashion, the potential for a valid process is increased. Moreover, as mentioned earlier, the question of replicability of decisions should be of special interest to model programs. It is suggested, therefore, that model programs make a special attempt to: 1) specify in unambiguous and objective terms exactly what each child assessment decision should accomplish; 2) determine the degree to

which those outcomes are actually achieved; and 3) test the replicability of decision by seeing whether different people, provided with the same information, would reach the same conclusion. If an assessment system appears to be leading to consistently appropriate decisions, then perhaps no further concern for system development is warranted. If problems in outcome or replicability are discovered, however, each step in the assessment process should be carefully examined.

STEPS IN THE ASSESSMENT PROCESS

Despite the fact that many decisions "seem obvious" or are made on the basis of "snap judgments", the process of reaching virtually any decision can be analyzed in terms of several discrete steps (Cooley and Lohnes, 1976; Anderson et al., 1975; Sax, 1974; Erickson, 1976; Stufflebeam et al., 1971): information gathering, organizing, analyzing, and choosing among alternatives. Each step in the process is important and (as with the weak link in a chain) a failure to carefully consider the best method for completing each step can result in unpredictable or inappropriate outcomes.

Information Gathering

Before any decision can be reached concerning a pupil's program, information must be gathered to describe the pupil and other factors which might influence the outcome of the decision. Because the need for precise and meaningful information is so obvious, this step in the decision-making process has received ample attention in the literature. Literally volumes of material have been

compiled discussing the types of information which should be sought and collected. Indeed, so much attention has been concentrated on this step in the process that educators are prone to reduce their concern for child assessment to the question of "which test should I use?" Even if one ignores the fact that such a question grossly oversimplifies the overall assessment process, the answer is still not simple. The selection of a test depends on the decision one wishes to make (a test which proves useful in child screening may prove worthless in developing an instructional plan), the types of pupils involved (methods for assessing object permanence in a sighted child are inappropriate for use with blind children), and a host of other situational factors (who is available to collect the information, when must the information be collected, how damaging will small errors in assesment results be, etc.) (Salvia & Ysseldyke, 1978). A formal test may not even be necessary or appropriate. Perhaps a questionnaire, a formal interview procedure, or even casual observation will serve as well or better in some situations.

There are no panaceas, no single type of information or procedure which will meet all needs. Most importantly, one must bear in mind that while the information collected can have a tremendous impact on the decisions made, information gathering is still only the first step in the complete decision-making process. If the information gathered is not amenable to systematic organization and analysis, or does not have a direct relationship to the available program options, all efforts will have been wasted. It will help, therefore, to get a more complete picture of the entire decision-making process before trying to reach any firm conclusions about the types of information which might be most useful.

Organizing the Information

"Raw data", as it comes from a test or observation, is frequently too detailed or complex to use directly. The important features of that information must be transformed and/or reduced to make it more easily understood and interpretable. For example, daily behavior counts and observation times might be translated into statements of "rate per minute" to place them in a common perspective. Those rates might then be charted or graphed in some way to make overall trends and daily fluctuations in performance more obvious (White & Haring, 1980). Similarly, the item scores from a standardized achievement test might be summarized according to means, percentiles or grade-level equivalents (Tallmadge, 1977), and related items on an attitude scale might be plotted together on a profile to make certain areas of special interest more apparent (Henerson, Morris & Fitz-Gibbon, 1978). Many commercially prepared assessment instruments include descriptions of how information can be transformed or displayed to make interpretation of results easier, but care should be taken to use only those procedures which are truly related to the decision one wishes to make. If, for example, the decision concerns acceptance of a child into some general program, percentile data which show the pupil's overall standing in major developmental areas may be useful. Those same summaries might obscure the specific item performances of greatest value to a teacher trying to decide where a pupil's instruction should begin.

Analyzing the Information

Analysis, for all intents and purposes, amounts to placing information into some meaningful perspective -- comparing it to some preset standard or level of expectancy. Without a comparison, most information will be meaningless. For example, if someone were to say a particular pupil got a "12", one would not know what it meant. Even if the "12" were defined as "12 digits per minute written correctly to answer math-facts", it would still be unclear whether that were good or bad. If, on the other hand, it was learned that most of the pupil's peers were able to complete at least 45 digits per minute on the same assessment probe, one might begin to reach a point where a meaningful decision could be made about the need for change in the pupil's program.

The standard of comparison used in an analysis is likely to have a tremendous impact on the eventual decision reached. In the example provided above, the pupil's peers were used as the standard, and the results of the comparison might suggest that a change in the program is advisable. A comparison might also have been made, however, between the pupil's current level of performance and his or her average performance last week. That comparison might indicate that the pupil is actually improving quickly and the program should be left alone. Neither comparison is necessarily better than the other. It depends solely on the decision one wishes to make and the assumptions upon which that decision will be formulated. If the decision concerns the integration of the pupil into a regular math program, and one makes the assumption that the pupil should be able to compete with his or her peers after integration, then the peer-referenced comparison makes sense. If, on the other hand, the decision concerns only the pupil's current program and its ability to facilitate continued progress, using previous performances as the standard of comparison is most defensible.

Choosing Among Alternatives

The final step in an assessment process is to make a choice among alternatives: Will the pupil be integrated into a regular math program or continued in a special program? Will a pupil's special program be changed or left intact? At least two factors will determine the outcome of this step and the overall success of the assessment system per se.

First, the alternatives available to the decision maker must be clearly defined (Stufflebeam et al., 1971). Information gathering, organization and analysis designed to clarify whether a pupil is making reasonable progress in his current program implies that an option exists to change that program. Is that really the case? In testing the effects of one set of decision rules, Liberty (Note 2) found that the most common excuse for ignoring a rule designed to prompt program changes was, "I didn't know what else I could try, so I left the program alone."

Second, one must consider whether personal judgment should be allowed in the decision-making process. Should the process lead to a definite, unequivable action (e.g., if the pupil's performance falls below a certain point, the program must be changed), or should the final choice of action be left to the discretion of the decision maker (e.g., if the pupil's performance falls below a certain point, the teacher is only advised to consider a program change)? Most assessment processes are really quite open-ended, even if they sound fairly straightforward. For example, a child study team may state that a decision to integrate a pupil into a regular program would not be made until he or she could complete work about as rapidly and as accurately as the normal peers in that program. Does that mean a level of performance equal to the mean peer performance? The median peer performance? Within one standard deviation of the mean? At a level equal to or higher than the 25th percentile of the peer group? Without getting more specific, there is obviously

a great deal of leeway in the final decision. That leeway might simply mean that all of the relevant variables cannot be specified in advance. It might also mean, however, that decisions will be capricious, postponed indefinitely, or inappropriate. Whenever possible, therefore, every effort should be made to provide specific guidelines for choosing among various program alternatives.

FACTORS AFFECTING EACH STEP IN THE ASSESSMENT PROCESS

When developing an assessment system, or refining one or more steps in an existing system, there are essentially five different factors to consider: the general outcome to be achieved, the person or persons involved, the timeline for completing the step(s), the specific procedures to be employed and, finally, the overall balance and efficiency of the process. As with the steps themselves, each factor represents a critical feature of the overall process and must be considered carefully.

What: The Overall Outcome

Each step in the decision-making process should produce some definite outcome: information gathering should result in a precise statement of one or more characteristics of the pupil or situational variables to be considered; organization should help to clarify the important elements and features of that information; analysis should place the information into some meaningful context and result in an appropriate

comparison with a standard or expectancy; and the last step in the assessment process should result in some action taken to influence the pupil's program. Before getting too specific about those outcomes, it is wise to begin with a more general statement of what one wants to accomplish. For example, before stating that information will be gathered using the WISC (and then trying to determine whether the WISC is an appropriate test for use with the pupil involved), it might be better to consider whether any sort of information concerning general cognitive functioning is desired. Similarly, before trying to determine whether a mean level or a median level of performance might be more appropriate for summarizing the data, the need for some sort of average (as opposed to individual item scores) should be considered. Many debates concerning details will be avoided if general issues are resolved first.

Who: The People Involved

The greatest single limiting factor in any decision-making process is likely to be the people involved: How many are there? Is communication a problem? How much time do they have? What expertise or experience do they have? Before those questions are resolved, serious consideration of specific procedures will be premature. Perhaps, for example, information is clearly needed concerning a pupil's overall development. If test "x" will provide the most precise and reliable measure of that development, one would certainly be tempted to use it in information-gathering. If, on the other hand, that test were so complicated that only specially trained personnel could administer it, and no such people were available, alternatives would have to be sought. Problems in experience and expertise are most likely to crop up when more than one person is involved in the process. Schematic profiles of hearing impairment expressed in decibels on a log chart of the auditory spectrum might be

fine for a meeting of audiologists, but in a meeting with teachers and parents some additional "organizational" or "analysis" efforts might be required to make the implications of the impairment clear. Even when expertise and communication per se are not problems, time factors may place severe restraints on the viability of various options. For example, teachers found it difficult to employ a 20-second decision rule in their classrooms (White, 1971), but had no difficulty in applying a similar rule which took only two or three seconds (Liberty, Note 3; White & Haring, 1980). The time factor in each case may seem rather small, but if the decision rule has to be applied to the four or five programs for each of 20 or 30 children every day, the difference between 20 seconds and two or three seconds becomes substantial.

When: Scheduling the Assessments

Aside from the time required to complete a given assessment process, timing in a broader sense is often critical -- exactly when and how often must assessment be conducted? For example, pupil performance on a standardized achievement test might prove useful for making decisions concerning referral to special programs, but if such tests are not given at roughly the same time of the year used to develop the original norms (usually sometime in October/November and/or April/May), the child's true standing with respect to the original group may be seriously under- or overestimated (Tallmadge, 1977). Similarly, decisions involving major curriculum changes are frequently most efficiently made at the end of a school quarter or year, while decisions involving the modification of instructional strategies may have to be made daily or weekly (White & Haring, 1976; Haring & Liberty, Note 4). Broad temporal constraints will have a tremendous influence on the usefulness and appropriateness of specific procedures and should be

considered early in the development of any assessment system.

How: The Specific Assessment Procedures

After the basic "what, who and when" of the assessment system have been defined, it should be possible to identify and select specific procedures for each step in the assessment process. The level of detail necessary will depend in large measure on the degree to which the procedures represent "standard practice" and are commonly known. For example, it might be sufficient to state that "teachers will test each pupil during the last two weeks of October using the Iowa Test of Basic Skills following the procedures as outlined in the test manual," but it would not be sufficient to say that "teachers will summarize the salient features of the pupil's social development." The best way to determine whether procedures have been adequately specified is to have several people attempt to implement the procedures and then to examine the results. Edwin (Note 5) has suggested that the steps taken by an individual during the assessment process be recorded and then analyzed by "experts" to identify which operations were definitely part of the prescribed procedures, which were "neutral" (not part of the prescribed procedures, but still acceptable) and which were definitely not allowed. If most of the prescribed operations were followed, it might be assumed that the process is relatively well defined. As a further test of that assumption, however, Edwin advised that the person whose behavior was recorded also rate the record. If that person accurately identifies which operations were part of the prescribed process and which were not, then the clarity with which the process was defined is further supported (even if the person chose not to employ those operations in all cases); but if the person involved identifies some operations as part of the prescribed procedures when they were not (or vice versa),

then it must be assumed that the process was not clearly defined or explained, even if the person just "happened" to follow the procedures relatively well.

Overall Balance and Efficiency

While it is important that the overall outcome of each step and the entire decision-making process be valid (accomplish what it is supposed to accomplish) and reliable (accomplish results in a consistent manner), it is also important that the assessment be efficient. At least two problems in efficiency are common.

First, the overall cost of the system may be so great that it cannot (or will not) be widely replicated. One program, for example, developed a pupil-tracking system which allowed teachers to make precise predictions about pupil progress and appropriate decisions concerning placement and programing. The system depended upon a complex computer program to realize its full benefit, however, and could not be translated into languages used by other computers for less than tens of thousands of dollars. Obviously, the usefulness of that system to other programs is severely limited.

A related, but more subtle problem can arise when there is an inherent imbalance in the cost, effort or sophistication of the various steps within a single assessment process. It is often said of computers, for example, "garbage in, garbage out" -- referring to the foolishness of developing sophisticated analytic procedures for information of questionable value. The reverse may also be true -- poor analyses may destroy the value of potentially good information. One model program actually used 41 different assessment instruments with every referred pupil. Aside from the possibility that a smaller number of instruments might be used with a pupil (thereby reducing the cost of the overall

system), there was a fundamental problem in using all the information collected. While the information-gathering step in the process was clearly defined, the organization and analysis steps were described as only, "getting all the testers together to discuss the results." As Holland (1980) points out, such a loosely defined situation is unlikely to produce replicable or meaningful results. If the information-gathering phase of decision making warranted 41 tests, the organization and analysis phases should have received equally extensive attention.

WHICH DECISIONS?

Educators are continuously making one decision or another which could have a direct impact on a pupil's program. Some of those decisions obviously deserve the support of a formal, well-defined assessment system, and some are best left to whatever informal system seems easiest at the time. For example, the Education for All Handicapped Children Act (PL 94-142) makes it clear that decisions concerning placement and overall program goals do warrant a formal assessment process, but no one is likely to care how a teacher decides whether the upholstery for a child's wheelchair will be blue or green. There are some decisions in the middle where the need for formal assessment is less certain. The following points might be considered when trying to decide.

The Law

Assessments required by state or federal law should certainly be conducted as systematically as possible. When examining legal requirements, however, it might be

wise to remember the distinction between determinations and decisions. Many laws only govern the types of information to be collected and, perhaps, the people who should be involved in any decision reached. For example, PL 94-142 states that an Individual Educational Plan must contain a description of how progress toward intermediate objectives will be measured. It does not state that anything has to be done with that information. While there is an implication that a program will be modified if the pupil does not appear to be making adequate progress, implications do not always reflect reality. In order to ensure that the intent of the law is fulfilled, it is frequently necessary to go beyond basic requirements and develop a system which leads directly to some meaningful point of decision.

Options

As mentioned earlier, one or more options need to exist before a decision can be made. If, for example, a particular program involves only one class and one teacher/therapist, no "placement decision" system will be required beyond the initial decision that the program as a whole would be appropriate for a pupil. With a program involving several service/placement options, a system which only formalizes the process of initial acceptance might not be enough. Similarly, a model which requires that instructional programs be conducted for a minimum of two weeks before any modifications are considered might not need an assessment system designed to prompt and facilitate daily program change decisions, but a program which emphasizes a rapidly changing approach to instruction might find a system for making daily decisions very useful.

It is best to think beyond the immediate range of the model program when considering options. Perhaps, for example, a particular model program offers only one

approach to physical therapy. That would obviate the need for a formal approach-selection assessment system. If other sites are likely to offer several options, however, it might be wise to delineate the factors which should be considered when selecting among those options (e.g., the relationship of various options to the overall model; pupil characteristics most often associated with success). Even if that selection process is never actually implemented at the original model program site, the outline of such a system would make it easier for other sites to integrate the procedure into their existing program.

Probable Impact

Decisions which have the highest probability of having a profound impact on a pupil's eventual success or failure should be given the greatest attention when developing assessment systems. Major program decisions, like transfering the pupil to a new program, come most quickly to mind and are the most commonly considered candidates for formal assessment systems. Care should be taken not to overlook the cummulative effects of smaller, more frequently made decisions, however. Formal assessment systems have proven of significant worth even when applied to daily classroom instructional decisions (Bohannon, Note 6; Mirkin, Note 7; White & Haring, 1980; Haring & Liberty, Note 4).

Relation to Model

Each model program is, presumably, based on some coherent, logical and well formulated set of assumptions or philosophy. Decisions which are critical to the maintenance of a program in accord with those assumptions or that philosophy should be formalized to

whatever extent possible. For example, a program based on the assumption that each minute of a child's day should be highly structured and carefully planned will probably need a relatively formal system to decide when each instructional program should be conducted. Without a formal system for making those decisions, it would be difficult to ensure that replication of the program be consistent with the program's assumptions. If a program were based on the assumption that children should be allowed to provide "their own structure," a formal system for deciding when programs should be run would be of no use or concern.

COMMON DECISIONS

Although the specific child assessment needs of a program will depend upon a number of different factors such as those outlined previously, there are a few assessment decisions which must be made by virtually every program. Those decisions are outlined below along with a very brief statement of what an assessment system should accomplish in each case. Space does not permit a complete analysis of the issues involved in developing each type of assessment system, but by considering the general factors outlined earlier in this chapter, it should be possible for a program to identify the necessary components of an appropriate system. Many of the more common concerns are also covered in other chapters in this book and additional overviews are available elsewhere (e.g., White, 1980a).

Which Specific Programs and Services Should be Offered to a Pupil?

Pupil placement and the arrangement of support services is considered by many to be the single most important set of decisions which can be made for a pupil. More laws, court rulings and regulations have centered on this aspect of child assessment than any other. At least two general areas of concern seem most common: referral and initial program selection or placement.

Should a Child Be Referred? The overall purpose of a child-find, screening and referral system is to identify those pupils who might be in need of special (or at least different) services or programs. The question is not so much whether a pupil is handicapped, but whether there is sufficient reason to suspect a problem that should be investigated more closely. There are two basic criteria for an effective and efficient child-find system: 1) all of the pupils who are referred should be referred (i.e., they meet the qualifications for acceptance into the program); and 2) all of the pupils who should be referred are referred (i.e., no pupil who would qualify for the program is overlooked).

Where Should the Pupil be Placed? The main objective at this stage in the assessment process is to decide which of several program options should be provided to best meet the needs of the pupil. That process is usually divided into three steps: 1) the verification that the pupil does, indeed, require some form of special education or related service (as opposed to the regular education program alone); 2) the identification of the specific needs which must be met; and 3) the selection of actual placements and services to meet those needs.

Those steps are usually called, respectively, "initial assessment," "diagnosis," and "placement in the least restrictive environment." Together they constitute the

foundation of IEP development as specified by PL 94-142. Two basic criteria can be applied to the validation of a placement assessment system: the pupil will 1) make progress in the placement selected, and 2) make more meaningful progress in the selected placement than in any other possible placement.

Exactly How Should Instruction or Therapy Proceed?

Once a pupil has been placed in a program and basic services have been arranged, many educators tend to think that the process of formal child assessment has been completed. For the teacher or therapist, however, the burden of making educational decisions has just begun. Choices must still be made concerning the development of initial instructional plans and the way in which those plans will be monitored.

How Should the Program Begin? Taking the broad goals and objectives established by the child study team during the development of the pupil's IEP, it is the purpose of this phase in the assessment process to aid the teacher or therapist in deciding exactly where and how instruction should begin. A goal might be established, for example, to increase a pupil's self-help skills by teaching him or her to dress. Decisions still need to be made concerning the specific skill to be taught first (e.g., shoe tying or buttoning), where in the sequence of instruction to begin (e.g., lacing or bow-making) and what types of cues and consequences to use (e.g., physical prompts vs. extra verbal cues; praise vs. small bits of food). If the system for making initial instructional decisions is working, two criteria will be met: First, the pupil will make progress during the initial phases of instruction on each program; and second, the pupil would not have made greater progress with any other instructional plan.

When Should an Instruction Program be Changed? Regardless of the care with which initial programs are designed and implemented, it is unlikely that any single instructional plan will remain effective or appropriate throughout the entire year. It is the purpose of this system, therefore, to monitor each pupil's performance to determine exactly when and how his or her programs should be modified. The criteria by which this step in the assessment process might be evaluated are essentially the same as those presented for developing initial instructional plans (i.e., reasonable rates of progress; greater progress than with other programs). Attention must now turn to assessment procedures which will allow the teacher to detect the need for changes in programs. Two questions in particular need to be answered: 1) are problems in previously effective programs being detected and remediated in a timely and efficient manner; and 2) is the pupil moving as rapidly through the curriculum as is possible?

Should the Pupil's Placement be Changed?

The decision to change a pupil's placement is essentially a reinvestigation of the question discussed earlier in this chapter concerning the pupil's initial acceptance into a program. It is presented here as a separate question only to emphasize three points:

First, the process of assessment is a never-ending cycle. Each part of the assessment process should lead directly into another -- screening into general needs assessment, general needs assessment into placement, placement into program development, program development into program refinement and program refinement back into general needs assessment. Each step in the sequence should be related to every other step, which raises the next point.

The first annual review of a pupil's program by the child study team should not be a simple replay of the initial IEP development meeting. At the very least, more experience will have been gained with the pupil and, hopefully, a great deal more information will be available concerning the conditions under which the pupil is able to learn and progress. Special arrangements must be made to organize and analyze that information in a systematic way.

Finally, while the first child study team may have been convened to decide whether the pupil should be placed in a special education program, all subsequent meetings should determine whether it is now possible to move the pupil into a less restrictive program or out of special education altogether. The criteria for making that move should certainly be as explicit as any established for an annual goal or intermediate objective, and while it may be necessary to reanalyze those criteria to determine whether they are still appropriate, the focus of all efforts should be to achieve those ends.

CONCLUSION

The role of child assessment in any educational program cannot be underestimated. For model programs, however, child assessment systems do even more than provide a means for reacting in a timely and appropriate manner to the individual needs of each pupil. They facilitate model replication by providing an explicit framework for making the decisions critical to the faithful implementation of the model.

In order to be assured of reliable and valid results in child assessment, several factors need to be considered: the general purpose of the assessment, the people who must be involved, timelines for completing each assessment,

and the specific procedures which will be used for gathering, organizing, analyzing and using information to make specific choices.

The specific decisions which will warrant formal child assessment systems in any given program will depend upon legal constraints or mandates, the existence of options, the probable impact of decisions on the success or failure of a pupil's program and the assumptions upon which the model was developed. In most programs, child assessment systems should be considered for deciding whether a pupil should be referred and accepted into special education, where the pupil should be placed, how instruction should begin, when and how instruction should be modified, and when and how the pupil's placement should be changed.

FOOTNOTES

[1] Much of the material presented in this chapter was drawn directly from a more extensive work: White, O. R. Child Assessment. In B. Wilcox & R. York (Eds.), Quality Education for the Severely Handicapped: The Federal Investment, Washington, D.C.: Department of Education, Office of Special Education, 1980b.

[2] There are, in fact, several differences between the terms "assessment" and "evaluation." The discussions in this chapter will reflect their common educational usages. For a more detailed discussion of the formal distinctions between them, the reader is encouraged to consult White (1980b).

REFERENCE NOTES

1. Alper, T. Personal communication, Summer, 1979.

2. Liberty, K. A. Personal communication, Summer, 1980.

3. Liberty, K. A. Decide for programs: Dynamics aims and data decisions. Working paper No. 56, Regional Resource Center for Handicapped Children. University of Oregon, 1975.

4. Haring, N. G. & Liberty, K. A. Final Report, Field Initiated Research Studies of Phases of Learning and Facilitating Instructional Events For the Severely Handicapped. A grant from the Bureau of Education of the Handicapped, Project No. 443CH60397A, Grant No. G007500593.

5. Edwin, T. Procedural Adequacy. A presentation made at the national conference of the Association of Behavior Analysts, Dearborn, MI, 1980.

6. Bohannon, R. Direct and daily measurement procedures in the identification and treatment of reading behaviors of children in special education. Unpublished doctoral dissertation, University of Washington, 1975.

7. Mirkin, P. K. A comparison of the effects of three formative evaluation strategies and contingent consequences of reading performance. Unpublished doctoral dissertation, University of Minnesota, 1978.

REFERENCE LIST

Anderson, S. B., Ball, S., & Murphy, R. T. Encyclopedia of educational evaluation. San Francisco: Jossey-Bass, 1975.

Cooley, W. W., & Lohnes, P. R. Evaluation research in education: Theory, principles and practices. New York: John Wiley and Sons, 1976.

English, H. B., & English, A. C. A comprehensive dictionary of psychological and psychoanalytical terms. New York: David McKay, 1958.

Erickson, M. L. Assessment and management of developmental changes in children. St. Louis: C. V. Mosby, 1976.

Henerson, M. E., Morris L. L., & Fitz-Gibbon, C. T. How to measure attitudes. Beverely Hills: Sage, 1978.

Holland, R. P. An analysis of the decision making process in special education. Exceptional Children, 1980, 46, 551-554.

Salvia, J., & Ysseldyke, J. E. Assessment in special and remedial education. Boston: Houghton Mifflin, 1978.

Sax, G. The use of standardized tests in evaluation. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley: McCutchan, 1974.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. Educational evaluation and decision-making. Itasca, IL: F. E. Peacock, 1971.

Tallmadge, G. K. The Joint Dissemination Review Panel Ideabook. Washington, DC: U.S. Department of Health, Education and Welfare, 1977.

White, O. R. The glossary of behavioral terminology. Champaign, IL: Research Press, 1971.

White, O. R. Adaptive performance objectives: Form versus function. In W. Sailor, B. Wilcox, & L. Brown (Eds.), Methods of instruction with severely handicapped students. Baltimore: Paul H. Brooks, 1980a.

White, O. R. Child Assessment. In B. Wilcox & R. York (Eds.), Quality education for the severely handicapped: The federal investment, Washington, DC: Department of Education, Office of Special Education, 1980b.

White, O. R., & Haring, N. G. Exceptional teaching: A multimedia training package. Columbus: Charles E. Merrill, 1976.

White, O. R., & Haring, N. G. Exceptional teaching, (2nd ed.). Columbus: Charles E. Merrill, 1980.

Selecting and Evaluating Educational Tests

Corrine A. McGuigan

> All actions are risks. The present is the moment of decision, and by the decision taken the yield of the past is gathered in and the meaning of the future is chosen. The meanings of past and future are enclosed and are waiting, as it were, to be unveiled by human decisions.
>
> Rudolf Bultmann

Any particular philosophy is developed as a result of experiences, insights, prejudices and judgments. The development of a philosophy of testing is no exception. An attitude toward testing is the direct result of experiences in selecting, giving, scoring and interpreting tests and in reviewing the consequences of each action. Periodically it is wise to reflect on experiences in an area such as testing and to analyze them as they have been integrated into professional activities.

This paper focuses on issues of test selection and evaluation as they relate to assessment. The intent is to provide an overview of selection and evaluation considerations and, thus, to help educators systematically

reflect on assessment practices and outcomes. Divided into two sections, the paper first presents a discussion of overriding considerations which for many professionals lie at the heart of the assessment dilemma; practical and technical considerations are discussed in the second part.

OVERRIDING CONSIDERATIONS

In Anatomy of an Illness, Norman Cousins writes, "Our experiences come at us in such profusion and from so many directions that they are never really sorted out, much less absorbed. The result is clutter and confusion." Ann Morrow Lindbergh, in Gift from the Sea, echoes the same perception when she says, "We muffle demands in distractions." Both authors provide a basis for reflection: have educators, in the search of technically sound assessment practices, busied themselves with practical and technical considerations to the point of clutter and confusion? Is it still possible to see tests and testing practices for what they are, not for what they ought to be? Is it, despite the deluge of literature, still possible to lay firm hold of one of the most basic principles of assessment -- that is, assessment activities should never unnecessarily restrict the future opportunities of learners? Amidst the calls for evidence of validity, reliability, and item appropriateness, it must be remembered that the function of each assessment activity is to provide information which will give students more opportunities to succeed, not fewer.

Two points are to be considered. First, the instruments and processes used to assess students mirror the goals held for them. Instruments composed of finely sequenced social skills, for example, may prove extremely useful to the practitioner who must both assess and teach social development skills.This very specificity, however, may

result in teaching some students personally useless skills. For others, it may result in curtailing instruction by recommending insufficient success criteria. Too, it must be remembered by those who select assessment instruments that highly sequenced ("tight") instruments tend to "outlive" more open-ended instruments or processes. Because specificity makes instruments easy to use, they may be used for a long time, and perhaps reflect attitudes and expectations of previous generations. This phenomenon puts the test selector in a difficult position: should the test selected be general, (open-ended) or tight (highly sequenced)? Should it be based on realistic expectations of what society allows handicapped persons to do now, or should it reflect a more hopeful future outlook? The answer seems to lie, appropriately, in a balance of each. Professionals must imagine a better future for handicapped persons. Consequently, their selection of instruments must not only reflect present expectations, but possible future opportunities as well.

The second overriding consideration focuses not on issues of test selection, but rather on issues of test interpretation. Tests and testing practices must never be used to escape professional responsibilities. Certain assessment practices have allowed some professionals to avoid or minimize responsibilities by claiming that results indicate that a student may be incapable of further growth. Other professionals have insisted that students with a particular I.Q. were not capable of pursuing a particular educational or social goal. These notions, fostered by particular philosophies of testing, restrict personal freedom and imply that certain persons have innate learning ceilings.

Underlying errors in test selection or in interpretation by even one professional perpetuate inappropriate practices resulting in unfair and often unjust programs for students. Those concerned with proper assessment practices must continually ask themselves, "What are questionable practices and why have so many been tolerated so easily for so long? Why is it that testing leads to labeling?

Labeling to categorization? Categorization to class placement, and class placement to a set curriculum"? Is it that even in the best of times, the phrase "it's only a test" has been too readily accepted? It is not only a test score and it is not only a label. Most educators or parents have rarely seen a test score changed once it was entered into a permanent file; fewer still have ever seen a student once labeled, "delabeled." Arguing that a labeled child has as much right to advance as any other is not only naive, it is a debasement of truth. Once a child has been categorized, almost every subsequent educational activity promotes permanent membership in that group. Once a child is labeled, change is always the struggle to convince some other party -- be it parent, teacher or social worker -that the skills deemed appropriate for a nonhandicapped student may indeed be appropriate for a particular handicapped person.

The notion that particular programs are appropriate for particular types of students is not really at debate here. Many times proper assessments do make it possible to select the most appropriate program. Great care, however, must be taken in making such judgments and in using instruments with highly specific foci. Likewise, if student performance surpasses test predictions or expectations, the students should not be viewed as overachievers. Such a subtle assumption fosters the belief that it is the test which is maximally correct and that students are subservient to test scores. Few would argue that this assumption is an unequivocally deplorable error. Nonetheless, it does occur and it is the business of those who make, give, market, interpret and use tests to see that such assumptions are erradicated. It is the business of educators to maintain a philosophy of testing which ensures that test results will lead to greater freedom for students, not less, and to making more appropriate decisions about the future activities of persons, not fewer.

Rudolf Bultman has said, "The present is the moment of decisions, and by the decisions taken, the yield of the past

is gathered in and the meaning of the future is chosen." It is the responsibility of those working with students to ensure that assessment practices -- which include both selection and use of particular tests as well as adherence to specific processes -- are reasonably well developed and ethically sound. The need is great and the need is urgent because it's really not "only a test." Assessment practices reflect particular curricula which reflect particular goals. Goals reflect hopes, and hopes reflect dreams of better lives.

A mature understanding of assessment is often the result of difficult personal and professional explorations. The fruition is seen in an educator's ability to find satisfactory answers to pressing questions such as: What is the function of assessment? What are the long- and short-term consequences of inaccurate procedures? How is a student's future affected by the process used and the instrument chosen? The solutions to these and other questions and the subsequent evolution of a philosophy are difficult, but the evolution does enable one to enter well prepared into the analysis of technical and practical considerations of process and tool selection.

TECHNICAL AND PRACTICAL CONSIDERATIONS

Those concerned with the testing of exceptional students are acutely aware of the broad range of assessment practices which beset the field. Over the years, assessment has taken many forms, from subjective teacher referrals to highly sophisticated, complex and time-consuming batteries administered by educators, psychologists, speech and language specialists, physicians and other professionals. Obviously, a manageable system of assessment lies somewhere between these extremes. While no single process, procedure or assessment tool can

be said definitively to be "the best" for all learners, certain factors are critical to a meaningful selection of an assessment process or instrument.

Factors to be considered in the development of a system or in the selection of an instrument can be divided into two categories: technical and practical. Technical considerations are generally restricted to items which ensure the integrity of a process or tool. Of course, technical considerations must always be of paramount concern to the researcher or clinician who can and should closely control not only the process of assessment or the instruments, but the testing environment as well. Adherence to rigid technical considerations makes it possible to produce the same, or same type of, results, given specified learner attributes. Classroom personnel, however, often work in much less controlled environments and have much less control over what is used, and when. While they must be sensitive to technical considerations as much as possible, they also must attend to a host of pragmatic issues such as administration time, test costs and available personnel. Practical considerations such as these result from a series of given factors, including specification of curriculum and teacher-student ratio; practical considerations deal less with item integrity and more with outcome use.

The division of assessment considerations is not intended to focus attention on one division over the other. Quite frankly, it is as inappropriate for the practitioner to operate without knowledge of technical attributes as it is for the researcher to employ strategies which have no relationship to classroom programming. The goals for professionals working in the area of assessment must be to identify their own roles and to prioritize considerations based on the nature and intended outcome of their work.

Technical Considerations

In The Research Process in Education (Fox, 1969) six major technical considerations are identified for the evaluation and subsequent selection of educational tests: 1) validity, 2) reliability, 3) sensitivity, 4) appropriateness, 5) objectivity, and 6) feasibility. These items form the core of elements most frequently included in technical considerations for test selection and evaluation (Ferguson, 1976; Hardyck & Petrinovich, 1969; Salvia & Ysseldyke, 1978).

Validity. Test validity refers to the extent to which a test measures what its authors or users claim it measures (Fox, 1969; Salvia & Ysseldyke, 1978). For example, if a test purports to assess independent spelling skills, the test should be constructed in such a manner that the student independently spells words when given oral cues. While it might seem odd to measure what might admittedly be a side effect of a skill, this often happens, especially in such academic areas as reading or phonics. Consider the number of times a test selector claims to assess a student's ability to distinguish vowel sounds by having a student match letters to pictures. The reality is that it is quite possible for a student to match the letter "a" to a picture of an apple without ever knowing how to make the sound /a/. Conclusions about a student's ability to perform a skill should never be based on evidence relating to skills taught incidentally.

In addition to the obvious face validity of tests (i.e., it looks like "this test item and accompanying directions" will directly assess "this skill"), three other validity variables are often cited in the literature: content validity, criterion-related validity and construct validity.

Briefly, content validity refers to the appropriateness of the items on the test, the completeness of the item sample and the level of mastery at which the content is assessed. The importance of each of these considerations

cannot be underestimated. Irrelevant or tangential items may result in findings which indicate that a student has not acquired a certain set of skills and an inadequate item pool or sample may greatly complicate this problem. Consider, for example, a sample of two-syllable words which includes "tractor," "combine" and "acre." While these words may be valid, their validity increases in rural states where they are frequent and necessary parts of a student's written and spoken language; their validity decreases, especially if the item sample is small, in more urban settings. A student may well know how to decode, but the foreign nature of the words once they are decoded may result in student error.

Criterion-related validity refers to the extent to which a student's success on a criterion-related measure can be predicted from that student's score on an assessment test or subtest. An example may best illustrate the concept of criterion-related validity:

> Mr. Michaels gave his class a series of subtests from a frequently used math assessment tool. After analyzing both group and individual results, Mr. Michaels attempted to identify "ability groups" -- placing students with "predicted" like math competencies in the same group. The results from the initial assessment enabled Mr. Michaels to predict that certain students had mastered certain math skills which he then proceeded to verify. As soon as the ability groups met, Mr. Michaels tested the reliability of the predictions by administering the criterion-related tests from the students' workbooks.

Inasmuch as the initial test predictions correlated with actual student performance on the criterion checks, the test can be said to have criterion-related validity; inasmuch as it failed to predict success accurately, either through over- or under-identification, the initial assessment instrument can be said to lack criterion-related validity.

Construct validity refers to the care with which a test author has formulated proper hypotheses regarding the appropriateness of test format and item selection and 2) tested those hypotheses in an empirical manner (usually through a series of field-based studies). The testing manual should provide a description of procedures used in designing the testing instrument, a clear explanation of field-test procedures and a sufficient discussion of field-test results. Both statistical significance and educational relevance should be mentioned in the latter discussion as well as specifics regarding population and settings. When there appears to be poor or questionable evidence that research hypotheses have been supported (or, in the case of the statistical or null hypotheses, rejected) then it is imperative that 1) sufficient explanation and justification be given for the construction of the test, such as it is, and 2) a discussion of procedural limitations (e.g., use with a large group of students) and restrictions (e.g., restriction to paper/pencil responses) be stated.

The complexity of test validity makes it an obviously difficult consideration. For example, not only may each aspect of validity vary in quality, but any aspect may vary in a kind of "pseudo-trueness" when given particular populations or testing settings (i.e., a test assessed as "highly valid" in one setting may be somewhat less valid in another). Such intricacy makes it difficult to identify any particular instrument as valid without qualification. Even those who have analyzed assessment tools cautiously find questionable validity in some of the most commonly used instruments on the market. Table 1, for instance, lists a number of frequently used tests which Salvia and Ysseldyke (1978) found to have questionable validity. Obviously, the fact that they have been identified as questionable does not make them useless; it does imply, however, that the test selector should be aware of such realities and make decisions regarding selection and use accordingly.

Reliability. Reliability is the second major technical consideration. It refers to the consistency with which an

TABLE 1

TESTS WITH QUESTIONABLE VALIDITY AND/OR INADEQUATE RELIABILITY DATA[1a]

Test	Questionable Validity	Incomplete Or Inaccurate Reliability Data
Arthur Adaptation of the Leiter International Performance Scale		X
Bender Visual Motor Gestalt Test	X	
California Achievement Test[b]	X	
Developmental Test of Visual-Motor Integration[b]	X	
Developmental Test of Visual Perception	X	X
Durrell Analysis of Reading Difficulty[b]	X	X
Full-Range Picture Vocabulary Test	X	X
Gates-MacGinitie Reading Tests	X	
Gates-MacGinitie Reading Diagnostic Tests[b]	X	X
Gilmore Oral Reading Test		X
Gray Oral Reading Test[b]	X	X

McGuigan

Henmon-Nelson Tests of Mental Ability	X	
Illinois Test of Psycholinguistic Abilities	X	X
Metropolitan Achievement Test[b]	X	
Purdue Perceptual-Motor Survey	X	
Stanford-Binet Intelligence Scale[b]	X	X
Wide Range Achievement Test	X	
Primary Mental Abilities Test		X
Quick Test		X

---

1 Salvia, J., & Ysseldyke, J.E. Assessment in Special and Remedial Education. Boston, Houghton Miffin Company, 1978.

a Validity is extremely limited for nearly all socioemotional tests.

b No validity data are included in the manuals for these tests.

assessment instrument or process produces the same results under similar circumstances. Of special concern to the educator is evidence that the instrument or process produces the same (or same kind of) results across settings, administrators or time periods.

To comprehend the importance of test consistency, one need only imagine the chaos which would occur if test results could be altered simply by changing the test administrator, or the time or setting in which the test was given. A well-conceived and well-constructed test will produce the same score, or nearly the same score, regardless of extraneous factors.

In Handbook in Research and Evaluation (Issac & Michaels, 1971), four categories of variables affecting reliability are discussed in detail: 1) lasting and general characteristics of the individual, 2) lasting and specific characteristics of the individual, 3) temporary and general characteristics of the individual, and 4) temporary and specific characteristics of the individual. In the first two categories, the following are identified as possible sources of variance in test reliability: general student readiness, test wiseness, self-confidence, knowledge of specific skills, habits and response to certain items. In the latter categories, such factors as health, fatigue, fluctuation in attention, memory and luck are all identified as sources of variance in test reliability. For the most part, these factors cannot be controlled, but by being aware of each, test givers can minimize the possibility of detrimental effects. It is, for example, possible to minimize detrimental effects of test readiness and responses to certain types of items by giving students a series of preparatory exercises. It is possible to help students overcome feelings of defeat caused by one or two incorrect responses by helping them understand that they are not expected to know answers to all problems.

Like test validity, it is often difficult to ensure test reliability to the degree one might like. In part, this is due to the number of factors which affect test reliability;

in part, it is due to the difficulty in gathering complete reliability data. Referring again to those tests which Salvia and Ysseldyke (1978) found to have incomplete or inaccurate reliability data (Table 1), one immediately recognizes a number of commonly used tests. That these tests have questionable reliability does not render them useless. Like instruments with questionable validity, it does mean that test evaluators and selectors should be cognizant of such weaknesses and place only as much credence in an instrument as appropriate. Instruments, however, such as the Developmental Test of Visual Perception, The Durrell Analysis of Reading Difficulty, or the I.T.P.A., which have both questionable validity or incomplete reliability data must be carefully scrutinized.

Sensitivity. A third technical consideration is test sensitivity. Test sensitivity refers to the degree that a test makes the kind of discriminations desired. For example, if a test is to screen students for vision handicaps, it must in fact be able to discriminate correctly those students with and without vision problems.

Test sensitivity is at least as important as test validity and reliability. Because the consequences of improper identification are so far-reaching and so directly affect the lives of many children and youth, test selectors must use tests with maximum sensitivity. The need for such precision is highlighted by the advent of innumerable legal actions taken on behalf of students who have been inappropriately identified as handicapped or by those who were not served because they were not identified as needing special services. Either error in identification is serious.

In evaluating a test for sensitivity, then, selectors must seek information which describes the levels of Type I (alpha) and Type II (beta) errors. This specific information should be found in the tester's manual or the publication guideline. Even though it might require both time and study to understand this information, it is important to do so because these statistics indicate the

probability that the test will erroneously identify someone as handicapped. It also indicates the probability of missing the desired discrimination. If the information is not available in publication manuals, the only other way to determine the actual sensitivity of a test is through the trial-and-error process of administering the instrument and carefully analyzing the results.

Appropriateness. Test appropriateness is the fourth technical consideration. The term refers to the correctness with which the test giver adheres to the directives of the testing guidelines. Assuming that test givers do not intentionally wish to violate test appropriateness by deviating from guidelines, it becomes the responsibility of the test authors to provide explicit directions regarding test administration and scoring. These directives should include exact specification of the population for which the test has been designed (e.g., elementary students, grades Third and Fourth), setting or settings in which the test is to be given (e.g., regular classroom, large group), identification of test administrator and scorer (e.g., certified psychologist) and other specifics such as time per subtest and readiness exercises. Anytime a test user deviates from the specifications of the testing manual, he or she risks the chance of using the test inappropriately and, therefore, inaccurately.

Other considerations regarding test appropriateness are more subtle and may or may not be mentioned in the testing manual, issues such as reading level and language appropriateness and the appropriateness of items to learning goals (items too often overlooked). Before a test is selected and given, it must be carefully evaluated for the relationship between the test items and subsequent interpretations. Perhaps the most common error of test givers is an erroneous inference based on test results (e.g., inferences about a child's home stability based on a line drawing, or social skill mastery based on a self-reporting inventory). Test evaluators should always insist that there be a direct and logical relationship between the

test item and inferences made as a result of particular responses. If this relationship is not obvious or does not appear logical, the test giver should be immediately suspect of the item appropriateness.

Objectivity. A fifth technical consideration is test objectivity. An objective test or testing practice is one in which there is only one correct response for each question asked. The truly objective test provides the test giver (as well as the test taker) with explicit testing directions which include scoring guidelines. The specificity of the guidelines allows the scorer to mark items without having to interpret answers. The more specific the scoring directives, the more objective the test item. The greater the opportunity for the scorer to interpret an answer, the greater the opportunity for scorer bias and thus, the greater the chance of error in being objective. It may be helpful, therefore, to remember that there is an inverse relationship between the number of possible correct responses and test objectivity; as the former increases, the latter decreases.

Feasibility. The sixth and final technical consideration is test feasibility, which is an often overlooked dimension of test selection and utilization. Feasibility, according to Fox (1969), refers to the cost of obtaining and administering the instrument and giving and scoring the test (e.g., special training cost for administration or scoring). Logically, test cost will increase with each special requirement (for example, a need for specialized personnel or computer information centers). Of course, cost will rise drastically if specialized instruments must be purchased for any part of administration or scoring. A more subtle fiscal consideration is the possibility for test reuse. An instrument which can be used only once has far less utility than one in which replacement sheets can be purchased without renewing the entire, initial investment.

Time is also a consideration of test feasibility. Certainly the adage "time is money" is as accurate in education as it is anywhere. Tests which absorb too much time -- that

is, the consequence does not match the investment -- are to be excluded from serious test consideration as surely as those which have low reliability, validity or objectivity. The most feasible, and therefore the most desirable, tests are those which can be administered in a timely cost-efficient manner.

Tests which can be administered by classroom teaching personnel are especially desirable for two reasons: 1) the hiring of ancillary personnel is not necessary to complete the assessment and, even more important, 2) the assessment information becomes immediately available to the person who is likely to need and use the information first -- the teacher.

Technical considerations provide test evaluators and selectors with often critical guidelines for interpreting the integrity of an instrument or process. Analysis of tests and processes using such considerations helps professionals locate the most technically sound and, therefore, the most desirable instruments and processes. It must be remembered, however, that even a test or process which adheres to each and every technical consideration may not be the best in a given situation. Analyses of tests and processes must go beyond technical considerations if they are to be used effectively in public and private school classrooms. It is necessary to examine in concert with the technical issues a series of practical issues. It is these practical concerns which are presented in the following section.

Practical Considerations

In addition to many technical considerations, the test selector/evaluator is called upon to be sensitive to issues which are of paramount concern to practitioners. Included in a category of considerations appropriately titled "practical" are such questions as, do the assessment

tools and processes 1) reflect local curriculum expectations? 2) make desired discriminations reflecting local, state and federal guidelines for exceptional child placement? 3) result in information which is easily understood by teachers, parents, students and other support staff? 4) include more than one opportunity for a student to indicate success or lack of mastery on a skill? 5) include more than one mode of collecting assessment information? and 6) result in information which is useful in setting annual goals and establishing meaningful classroom programs?

As attested to by those who have worked in regular and special classrooms, the need to respond to practical considerations is clear. Throughout the years, educators, administrators, psychologists, and other professionals have been frustrated by time-consuming, complex and often unnecessary assessment practices to the point of "clutter and confusion." Fortunately, this very frustration has led to rapid change in many assessment practices. It has increased awareness of the need of practicality by those who prepare and market tests. Today, not only do classroom personnel request information but they expect it.

Curriculum Issues. It seems only reasonable that students be assessed on information or skills that they have 1) acquired in the past or 2) will acquire in the future. Logically, this implies that a high correlation exists between test items and curriculum items. The overall usefulness of a test is the measure and strength of this association. As the relationship between assessment items and curriculum increases, so increases the usefulness of the test; as the association decreases, so too decreases test usefulness.

The importance of the relationship between test content and curriculum leaves open two options for test selectors or evaluators: they can identify local curriculum components at each grade level and search out appropriate tests which reflect that curriculum, or they

can identify a rather substantial test and build a curriculum around it. (Obviously, the latter seems almost bizarre. One must remember, however, that in newly established districts or schools the concept may not be totally inappropriate.)

The process of identifying curriculum elements by grade is the most common method of identifying curriculum components. The task is achieved in a number of ways, some of which are certainly more time consuming than others. For example, district curricula by grade level can be specified by: 1) a curriculum committee in the district, 2) the department heads of curriculum areas (e.g., English, History), 3) the school board members or their appointees, 4) the principal in a given school, 5) a volunteer group of teachers, or 6) a committee of educators, parents and other support groups.

The important fact here is not how the information is gathered, but that it is gathered accurately. Once a third grade teacher, for instance, feels secure in identifying "money skills" as something to be taught, the test selector can be certain that a search for a test reflecting that skill will not be in vain. Likewise, the same information becomes useful to test selectors as they consider items for review in the fourth grade assessment program.

The process of identifying curricula and relevant tests is notably difficult. It would not be unusual for the process to take nearly a year to complete with teams of professionals working arduously to modify existing tests (or subtests), changing responses on certain item pools, deciding which norm-referenced test will best reflect the local curriculum, and so on. Unfortunately, the only alternative to this time-consuming endeavor is to approach testing in a random manner with no or little regard for the existing curriculum. The consequence of such an alternative is devastating, however, and should not be considered as a probable alternative for long, if at all.

Test Issues. That a test, series of tests, or testing practices enable professionals to make desired discriminations for appropriate child placement is the second practical issue to be considered. Not only must the selected test 1) have a low standard of error (that is, it must correctly identify handicapped children as handicapped and nonhandicapped children as nonhandicapped) and 2) comply with state assessment requirements, but it also must 3) make accurate discriminations at levels as close as possible to the standards set for local districts or schools. (The problem of the standard of error was discussed previously as a technical consideration and so is not discussed in detail here.)

Standards for exceptional child placement vary from state to state and, sometimes, from district to district. A few general considerations do pertain to most: Does the instrument reflect the type of modality assessment suggested or required (e.g., visual, auditory)? Does the instrument and process reflect the areas to be assessed as suggested or required (e.g., psychological, developmental, vocational)? Does the process include the necessary types of required tests (e.g., standardized, norm-referenced)? And lastly, does the instrument or process result in a level of identification commensurate with funding allocation standards?

It may seem somewhat insensitive to bring the issue of money into assessment considerations at all. The reality is, however, that only limited monies are available at this time to serve exceptional children. This means that assessments must accurately identify those students most in need of services while remaining aware of the needs of "high risk" students. Recognizing the reality of limited financial support for delivery of services to exceptional students leaves local personnel with important personal and professional decisions. They can select instruments which will identify even the most minute learning problems or even the slightest physically handicapping condition. Unfortunately, this approach may lead to

"over-identification," (i.e., the identification of a percentage of students "outside" the recommended levels of funding by category). They may select instruments which are costly and require specialized personnel and/or equipment, but which have remarkably high reliability and validity, and, therefore, result in levels of identification commensurate with those suggested by the State Department of Education (e.g., 2% to 5% of the children in the district have hearing impairments). Or, they may select instruments and processes which are efficient and can be administered by classroom personnel, but which screen only those students with the most obvious handicaps, therefore resulting in possible underidentification.

Ideally, it may seem wise to use the most sensitive instruments and processes available. Without financial resources to actually deliver services to students identified, however, it may not be the best practical decision. The consequence demands, therefore, that test selectors and evaluators follow a careful plan of investigation and weigh the advantages and disadvantages of any particular approach, given local and state requirements and restrictions.

Clarity of Assessment Information. A third practical consideration highlights the need for clear, concise and understandable assessment information. No longer do parents or educators care to wait patiently as clinicians ramble through technical jargon which finally results in a label for a student. The advent of parent involvement and subsequent legislation has brought about not only the need, but the requirement, that information be reported in a manner understandable to persons from varied disciplines. Additionally, the information must be immediately educationally relevant and must include discussion not only of weaknesses but of strengths as well. In short, no matter how "good" a test is thought to be, if its results cannot be interpreted in a meaningful way to nonspecialized personnel, then the instrument or process is itself suspect.

Test Modalities and Item Pools. The fourth and fifth considerations require that assessment findings have included 1) more than one opportunity for a student to indicate success or lack of mastery on a skill and 2) more than one mode of collecting the data. Both these considerations are critical for the same reason: incomplete information may lead to inaccurate findings, inaccurate findings to erroneous conclusions, erroneous conclusions to false labels and false labels to inappropriate class placements. Through careful review of the adequacy of test samples and appropriate response modes, selectors and evaluators can diminish the possibility of such errors. Careful examination of testing modalities (i.e., auditory, visual) can identify instruments which test for desired student discriminations and which test for them in the modality most used in the classroom or the learner's living/social environment. Examination of the quantity in the item pool indicates to the selector whether sufficient opportunities are presented per skill to indicate learned or unlearned information versus correct or error responses by chance. Item pools which include at least three opportunities on any skill item (e.g., words with short /a/ if the concept of short /a/ is being tested) can serve as a minimum standard.

That assessment, when possible, reflects more than one mode of data collection is important. Consider the differences in findings from an assessment which was only a parent self-report check list versus an alternate system which includes the parent self-report plus direct observation in the classroom and home. Quite obviously, the findings will vary, as well as the quality of documentation. When reasonable, then, total assessment should include direct assessment by 1) observation in relevant settings, 2) assessment of selected skills through paper/pencil test or direct cues to a learner and 3) systems such as interviews and self-reports. Perhaps the only caveat in this area is that the assessment process is limited in its scope by time, money and personnel constraints.

Usefulness of Assessment Information. This practical consideration is perhaps the most important: it is of paramount concern that any information collected during the process of assessment have direct use in helping professionals and parents identify realistic goals and meaningful classroom programs. Test evaluators and selectors must ensure that the information they collect is sufficient to this end. Implied in this responsibility is the need to collect reasonable amounts of information at various levels of specificity.

Levels or stages of assessment can, perhaps, be most graphically described by employing the image of a pyramid. As Figure 1 depicts, the base of the pyramid represents the most basic and general types of assessment data. Information gathered at this level provides data on how a group of children (and sometimes individuals within groups) perform in relation to national standards. Information gathered at this stage -- usually through norm-referenced, standardized testing -- is most useful in analyzing how specific groups compare in performance on a wide range of skills. Often, analysis of data at this stage can provide early information about students who may be candidates for further, more detailed assessments.

The second stage of the pyramid represents data gathered in specific skill areas such as math, reading or social/emotional development. Data gathered here should provide specific information on skill strengths and weaknesses in a given curriculum area and should result in assessing a student's performance in relationship to peers.

Stage 3 level assessment, the most specific, implies that further assessments must be conducted to obtain meaningful information for establishing classroom programs. At this stage, assessment personnel must be able to obtain information not only about specific skill strengths in an area such as math, but about specific skill mastery levels on certain skills (e.g., three times tables).

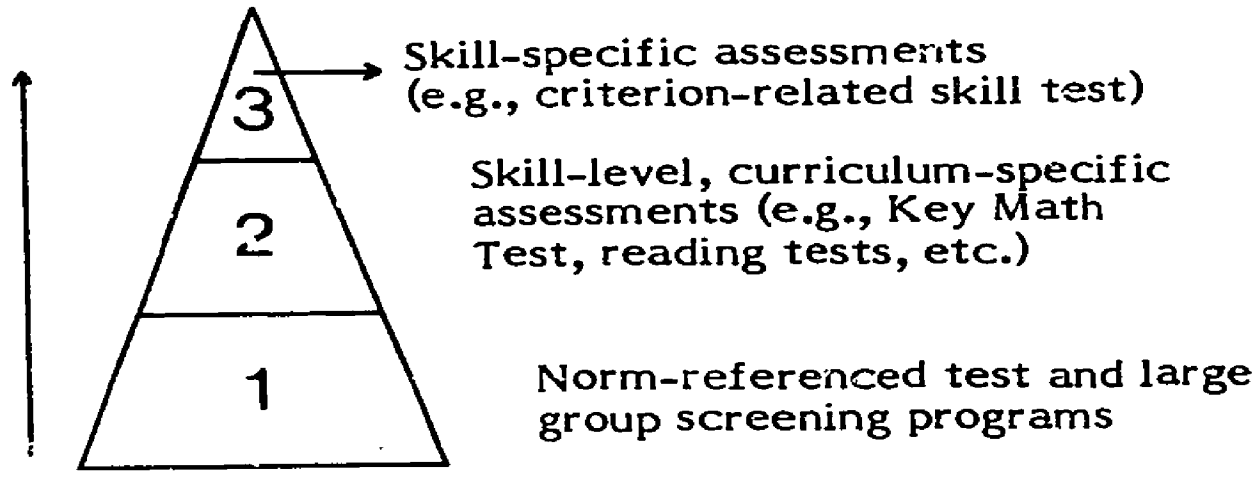


Figure 1: Stages of Assessment

Assessment is only completed well when it has truly touched in each area of the pyramid. Assessments which include only Stage 1 provide no information for programming; assessments which originate at Stage 3 may waste too much time on specific skill assessment when more appropriate data could be collected at Stage 2.

Sufficient information at each stage enables educators, parents, students and support personnel to have available information indicating student skills in relationship to national standards, peer standards and individual standards set for specific classroom materials. Information gathered at Stage 1, for example, helps identify students most in need of support services; Stage 2 makes more specific information available; Stage 3 enables programming committees or child-study teams to identify annual goals and establish meaningful classroom programs.

Given the scope and importance of each practical consideration, how are test selectors and evaluators to design and implement quality assessment programs effectively? The answer is certainly not easy, nor will it come quickly to personnel who must search out and identify the quality existing programs, the specific content of the assessment program, such as it is, and the philosophies underlying it. The best one can hope to do is to be keenly aware of practical considerations and to deal with them as best as possible. When negotiation must occur, then it should occur. When anything less than the ideal is settled upon, however, an understanding of the advantages and disadvantages of such decisions should be firmly rooted in everyone's mind.

CONCLUSION

> We all have to take chances in life. Humankind would be vastly poorer had it not been for persons who were willing to take risks against the longest odds. Our problem is how to remain properly venturesome and experimental without making fools of ourselves.
>
> Bernard Baruch

There is no black and white in educational testing, not even in test or process selection. There is no scoring procedure by which the benefits and limitations of tests, test scores and testing practices can be compared. Even the oft-touted objectivity of a test is, paradoxically, far from unambiguous. The fact is that even a test which conforms to all the requirements of instrument selection may not be, for a particular student, an objective or valid test.

Where does this leave educators who are searching for the best assessment practices? Frankly, it leaves them, as Baruch says, in the position of being "venturesome and experimental without making fools of themselves." This position is rather tenuous when the stakes are the lives and futures of individuals.

The challenge for educators becomes to select the best possible assessment systems using the best possible instruments to obtain the most informative, nonprejudical information. The challenge becomes to educate oneself, to reflect on a personal and professional philosophy of testing, and to act boldly according to one's beliefs and judgments.

REFERENCE LIST

Baruch, B. American industry. New York: Prentice Hall, 1941.

Bultmann, R. Theology (Vol. 1). Translated by Kendrick Grobel. New York: Charles Scribner's Sons, 1951.

Cousins, N. Anatomy of an illness. New York: W. W. Norton, 1979.

Ferguson, G. A. Statistical analysis in psychology and education. New York: McGraw-Hill, 1976.

Fox, D. J. The research process in education. New York: Holt, Rinehart and Winston, 1969.

Hardyck, C. D. & Petrinovich, L. F. Introduction to statistics for the behavioral sciences. Philadelphia: W. P. Saunders, 1969.

Lindbergh, A. M. Gift from the sea. New York: Random House, 1955.

Salvia, J. & Ysseldyke, J. E. Assessment in special and remedial education. Boston: Houghton Mifflin, 1978.

Assessing Social and Emotional Problems

Bob Algozzine[1]

It is quite common for lists of characteristics of exceptional children to include references to social and emotional problems. For example, most textbooks on learning disabilities and mental retardation have sections in which the social-emotional problems of these children are described and discussed. In fact, a separate category exists primarily for children with such problems; that is, emotionally handicapped, behavior disorders, emotionally disturbed, or similar synonyms are used to label these youngsters and define their group.

NATURE OF THE PROBLEM

In the state of Florida, the emotionally handicapped (EH) child is one who "after receiving supportive educational assistance and counseling available to all students, still exhibits persistent and consistent severe to very severe behavioral disabilities which consequently disrupt his/her

own learning process. This is the student whose inability to achieve adequate academic progress or satisfactory interpersonal relationships can not be attributed primarily to physical, sensory, or intellectual deficits" (Algozzine, Schmid & Connors, 1978). Other similar definitions exist for children with behavior problems (Epstein, Cullinan, & Sabatino, 1977; Kauffman, 1977); it should be obvious that a vague generality permeates the state-of-the-art in identification practices which derive from such definitions.

To escape vociferous criticism with regard to the nebulous nature of disorders characterized by social and emotional problems, some states have attempted to "tighten up" the criteria for eligibility for services/assistance; Florida's definition offers some advantage here, in that it contains clauses for which operational criteria may be written. For example, the various levels of alternate placements may be delimited as part of screening and/or identification procedures; similarly, criteria for the "persistent and consistent severe to very severe" distinctions may be developed. Kauffman (1977) discussed a variety of problems in definitions of "emotionally disturbed children or children with behavior disorders"; his listing included measurement problems as foremost. While operational capabilities do not solve the measurement problem, they tend to make the identification process appear more objective than is evident when definitions do not possess them.

Knowing the extent to which a child exhibits "behavioral disabilities" is only half the problem of assessing social and emotional problems (and it may not be that much). It is still necessary to know the actual behaviors which are the source of concern.

The Cataloging Procedures

Traditional approaches to the development of these behavioral characteristics have taken the "undistributed middle term" approach to identifying problem behaviors. It goes something like this: Surveys of individuals with known histories of behavior problems are completed and listings of their characteristics are prepared. Statistical procedures are then applied to these listings and rating scales are born from the results. The logic of this approach has been previously discussed (see Algozzine, this volume for a review); the following sequence of statements is an example.

Crazy people (A) do these behaviors (B).
Ratings of you (C) suggest that you do the behaviors (B).
We think you (C) are crazy (A).

An alternative approach is to view the relationship between the behaviors and the outcomes initially and then to identify the extent to which the presence or absence of behaviors is relevant to treatment. For example, the argument might go like this:

Certain behaviors (A) interfere with learning (B).
We observe in you (C) those same behaviors (A).
We predict in you (C) problems in learning (B).

The focus of this argument is that behaviors are the source of the problem and, therefore, are the source of the intervention. All that is necessary is to catalog the behaviors which interfere with learning.

Only one hitch appears in this logically sound argument: that is, behaviors may interfere with learning for a variety of reasons, some of which do not only depend on the individual exhibiting the behavior. Algozzine (1979) has shown that behaviors on the Behavior Problem Checklist (Quay & Peterson, 1975) are "disturbing" to teachers in working with school children. This suggests

that part of the problem of the child is the fact that what he or she does is bothersome to a teacher. Similarly, such research casts doubt on the utility of teacher ratings as a measure of social and emotional behaviors; the rating of the extent of occurence (e.g., always, sometimes, never) may be as much a function of the nature of the behavior relative to a teacher's tolerance as it is the actual frequency of the behavior in the child's repertoire. The recommended practice, then, is to use rating scales as a last resort in identification, intervention, and/or evaluation of special children and their educational programs; they may be biased and insensitive indicators of change. An alternative approach is to measure the extent of occurence of a behavior rather than to take someone's word for its existence.

Social and Emotional Behaviors

A variety of behaviors has been shown to interfere with productive interpersonal relationships and/or to coexist with behavior problems (Algozzine, 1979; Cartledge & Milburn, 1978; Cobb, 1972). Cobb and Hops (1973) identified a set of behaviors which interfered with academic success and which they termed "survival skills". They then taught low-achieving first graders and reported that when social survival skills improved, so did achievement levels. Other similar results have been obtained (Hops & Cobb, 1973, 1974; Walker & Hops, 1976). A list of interfering behaviors is presented in Table 1; the staunch behaviorists will quickly suggest that "low self-concept" is not a behavior. They are correct; however, it is a relatively simple task to make it observable (so what difference does it make??).

TABLE 1

SOME POSSIBLE INTERFERING BEHAVIORS

Behavior	Useful Abbreviation
Emotional Varieties	
Low frustration tolerance	LFT
Low self-concept	LSC
Negative over-reaction	NOR
Limited range of emotional reactions	LRE
Impatience	IMP
Anxiety	ANX
Temper Tantrums	TNT
Paranoid reaction	PAR
Social Varieties	
Task avoidance	TAV
Disruptiveness	DRP
Non-attention	NAT
Irrelevant activities	IRA
Slowness in work	SIW
Achievement anxiety	AAX
Low management skills	LMS
Disrespect/defiance	DDF
Limited comprehension	LCM
Low academic achievement	LAA
General social withdrawal	GSW

Response Definition

For the most part, scientists (behavioral as well as more general social ones) engage in an imprecise field of study. For example, many of the special education related disabilities, disorders, and dysfunctions are arbitrarily defined and identified; in fact, many problems exist because professionals say they exist. Similarly, the units of measurement within the social sciences are often subjectively determined. It is important here to differentiate the unit being measured (i.e., the behavior) and the measurement unit (i.e., the count or occurrence of the behavior). The measurement units in assessing social and emotional behaviors vary in definition and level of precision (in fact, the definition and level of precision are at the discretion of the observer). For example, it is possible to measure various kinds of occurrences on an hourly, daily, and/or weekly basis. The precision obtained will be a function of the nature of the behavior and how much of that behavior occurs during the measurement time.

Similarly, the units being measured are subject to various degrees of definitional precision. Definitions of behavior may be differentiated along a "continuum of inference". Those behaviors which are at the low end of the continuum (i.e., low inference behaviors) require less interpretation for identification. For example, nose picking or hand raising may be thought of as low inference behaviors; an occurrence may be interpreted as a discrete unit and, in that sense, is more precise. High inference behaviors, on the other hand, are not observable as discrete units but are inferred through the observation of associated or representative low inference behaviors. For example, low self-concept is a high inference behavior; it is observable only through identification and tabulation of other reference behaviors thought to be representative of it (e.g., hand raising in response to open questions, negative self-statements, and/or volunteering answers).

Given that the best alternative to assessment of social and emotional behaviors is one in which actual observations are used rather than (or in addition to) rating scales, operational definitions for the social and emotional behaviors of concern must be developed. Some examples of possible operational definitions of selected interfering behaviors are presented in Table 2. The extent to which a common definition of a behavior exists, the more likely agreement will occur in identifying or labeling the behaviors operationally defined in Table 2 or any other behaviors which interfere with productive interpersonal relationships.

The task then in assessing social and emotional behavior is to establish a response class, that is, a defined, identifiable behavior or group of behaviors. The class may be a low or high inference unit. The choice of each should be measured by the direct relevance it has for assessment. If the task is to identify and remediate "nose-pickers," then the response class may be a low inference one; if it is to identify and remediate children with "low self-concepts," the class will be a high inference one. When assessment for intervention effectiveness is the source of concern, then the same rules apply. If the purpose of assessment is to evaluate interventions for nose picking, then the low inference units are observed and so on.

AN OBSERVATION SCALE

Assessment of social and emotional behaviors boils down to definition of response classes and observation of reference units for the classes. To facilitate assessment, a simple observation scale is all that is necessary; an example is presented in Figure 1.

TABLE 2

POSSIBLE OPERATIONAL

DEFINITIONS FOR SELECTED BEHAVIORS

Anxiety (ANX). Given a situation in which performance is requested and/or expected, the child persistently and consistently engages in activities and responses which suggest that an unusual amount of apprehension or concern is associated with that performance.

General Social Withdrawal (GSW). Given a situation or activity which involves interaction with others, the child persistently and consistently responds with statements and actions which reduce the likelihood of participation.

Irrelevant Activities (IRA). Given an opportunity to complete an activity or task, the child persistently and consistently responds with statements or actions directed toward productive efforts involving other activities besides the one at hand resulting in its postponement.

Low Self-Concept (LSC). Given an activity in which personal performance is expected, the child persistently and consistently responds with statements and/or actions which reflect anticipated failure or actual failure due to perceived personal inadequacies.

Task Avoidance (TAV). Given an opportunity to complete an activity or task, the child persistently and consistently responds with statements or actions of a non-specific nature relative to the task at hand but which do not result in task completion.

Low Frustration Tolerance (LFT). Given an activity which results in frustration but which is clearly within the child's response capabilities, the child persistently and consistently responds with statements or actions which reflect reduced likelihood of task completion.

Low Management Skills (LMS). Given an opportunity to participate in an activity or task, the child persistently and consistently responds with statements and actions which reflect limited awareness of expected social behaviors and/or self-control.

FIGURE 1: SAMPLE OBSERVATION FORM

Student _____ Date _____

Observer _____ Purpose _____

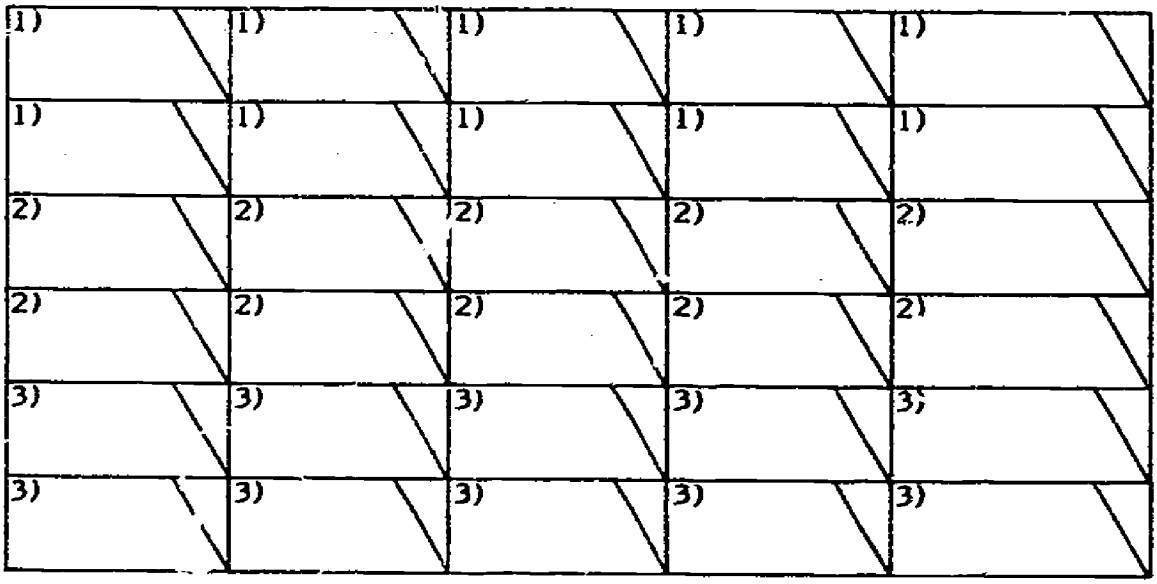
Response class of interest: _____

Reference Units:
(indicate observable behaviors)

1) _____
2) _____
3) _____

Measurement Unit:
(circle one)

frequency duration
(other) _____



Intervention Planning

1) _____

2) _____

3) _____

4) _____

_____

Background

The observation form includes information regarding the student being observed, the observer, the date(s) on which the form is utilized and the purpose of the observation (e.g., referral, identification, intervention, follow-up). The response class of interest is also identified on the form. Most high inference response classes which interfere with productive interpersonal relationships are not directly observable as discrete units; however, they are recognized by other representative behaviors. Many teachers identify these problems by the individual behaviors which have come to be associated as reference units for them. The observation form includes information about general response classes of interest (e.g., low self-concept, temper tantrums, etc.) as well as observable reference units (e.g., negative statements about one's own abilities, crying, etc.). Teachers reference response classes with different behaviors; similarly, children demonstrate response classes in different ways. For this reason, the form remains open-ended. The first task in any observation, then, is to identify the response class of interest and the target reference behaviors.

It is next necessary to choose a unit of measurement through which the observable reference units will be tallied. Many alternatives are available; however, for most purposes, frequency or duration is sufficient. Frequency should be used when the number of times a reference unit occurs during a period of time is of interest; duration should be used when the length of the occurrence is of interest rather than a simple count of it. Any measurement units other than frequency or duration should be indicated in the space provided.

How to Use the Scale

To use the form, an observer indicates the time period of the observation (e.g., minute, hour, day, 10 minutes, etc.) in the triangular right-hand corner of the appropriately numbered box (i.e., number = reference unit). Then the measurement units during that time period are recorded in the remainder of the box. These data may then be presented in other forms as needed (i.e., graphs, tables, etc.). The form also includes a section in which a teacher may develop some intervention plans based upon the result of the observation. This section can become the basis for a complete individual education program for each child and/or the basis for evaluative judgments regarding program effectiveness.

THE SOLUTION?

The state-of-the-art in assessment of social and emotional problems has been that rating scales with varying degrees of sophistication have been developed as a measure of the extent of the occurrence of selected "behaviors"; in fact, hundreds of such checklists exist. Ratings obtained from various professionals using these scales have been thought to be representative measures of social and emotional problems. A variety of issues relative to this form of assessment have been identified. An alternate approach seems to be one in which observations of selected behaviors are made. The task of assessing the extent to which (the measurement problem) specific social and emotional behaviors (the definition or identification problem) exist prior to, during, and after intervention may be facilitated by the development of response class definitions which include reference units of high or low inference. The use of the proposed

observation scale may be beneficial to professionals engaging in identification and remediation of relatively imprecise areas of behavior.

FOOTNOTE

[1]Bob Algozzine is also affiliated with the University of Minnesota's Institute for Research on Learning Disabilities.

REFERENCE LIST

Algozzine, B. The disturbing child: A validation report. (Research Report #8) Minneapolis: Institute for Research on Learning Disabilities, 1979.

Algozzine, B., Schmid, R. & Connors, B. Toward an acceptable definition of emotional disturbance. Behavior Disorders, 1978, 4, 48-52.

Cartledge, G., & Milburn, J. F. The case for teaching social skills in the classroom: A review. Review of Educational Research, 1978, 48, 133-156.

Cobb, J. A. Relationship of discrete classroom behaviors to fourth-grade achievement. Journal of Educational Psychology, 1972, 63, 74-80.

Cobb, J. A., & Hops, H. Effects of academic survival skill training on low achieving first graders. The Journal of Educational Research, 1973, 67, 108-113.

Epstein, M. H., Cullinan, D., & Sabatino, D. A. State definitions of behavior disorders. The Journal of Special Education, 1977, 11, 417-425.

Hops, H., & Cobb, J. A. Survival behaviors in the educational setting: Their implications for research and intervention. In L. A. Hammerlynk, L. C. Handy, & E. J. Mash (Eds.), Behavior change. Champaign: Research Press, 1973.

Hops, H., & Cobb, J. A. Initial investigations into academic survival skill training, direct instruction, and first-grade achievement. Journal of Educational Psychology, 1974, 55, 548-553.

Kauffman, J. Characteristics of children's behavior disorders. Columbus: Charles E. Merrilll, 1977.

Quay, H., & Peterson, D. Manual for the behavior problem checklist. Mimeographed, 1975.

Walker, H. M., & Hops, H. Increasing academic achievement by reinforcing direct academic performance and/or facilitative nonacademic responses. Journal of Educational Psychology, 1976, 68, 218-225.

Analysis and Use of Performance Data

Corrine A. McGuigan

Almost without exception, educators strive to provide students with the most effective educational programs possible. In so doing, many have found that the systematic collection and use of data on an ongoing basis provides valuable information about: 1) student performance on given tasks, 2) the effectiveness of instructional strategies or materials, and 3) the amount of information likely to be acquired in a specific time. Especially for educators striving to ameliorate or lessen handicaps, information made available through consistently monitoring performance is viewed as simply irresistible. For these educators the collection of data is seen not as obligatory, but rather as professionally desirable.

Such a positive association between measurement and instructional outcomes has not always existed. In the mid-1960s, when ongoing (or daily) data collection first came into vogue, it greatly divided educators and forced professionals with differing philosophies into two camps: those who viewed data collection as the modus vivendi of the field and those who valued it less than an electric typewriter without an energy source. Fortunately, such

divisions have healed themselves through increased understanding on the parts of many educators representing a wide range of views, competencies and philosophies. Today, most educators concur that measurement, kept in proper perspective, is critical to the implementation of effective instructional programs.

Knowing how (and how far) an understanding of data systems has evolved helps focus attention on present and future practices. Will even more realistic, efficient and understandable systems develop as a result of the continued work of practitioners? The answer appears to be a hopeful "yes." As educators devise and share practices, the knowledge base will expand. In turn, many more educators will be introduced to the advantages of data collection and the role it plays in effective instruction. It is the intent of this paper to introduce concepts of monitoring that will make instructional time richer, and planning for it easier.

The focus of this paper is not the collection of data per se, but the use of data. The underlying philosophy is founded in the simple belief that the mere collection of data is insignificant to students and teachers; it is the use of data to make decisions about the type, quantity or quality of instruction which makes the collection of data so compelling an activity and such an integral part of instruction.

Elements of this discussion are offered in three sections: 1) the establishment of aims, 2) the identification of program guidelines and decision rules, and 3) the analysis of data patterns to make appropriate educational or motivational interventions.

ESTABLISHING AIMS

The first step in utilizing data effectively is to identify an end or desired goal for the behavior being monitored. Only when such goals are established do individual data lend themselves to meaningful interpretation. For example, scores of "7" and "8" are meaningless, because from the scores alone, it is impossible to determine whether performance is "good" or "bad." The same scores become meaningful, however, wher a goal of "10" is stated. Aims provide the reference point for the appropriate interpretation of data as well as specifying for the student the precise expectations for acceptable, terminal behavior. In short, goals or desired aims (aka: "aims," "behavior aims") stated in terms of both rate and date indicate where one wants to go and how long one has to get there.

The establishment of aims is not always easy. It is a process which involves a number of steps beginning with 1) the specification of the behavior to be monitored in specific, observable terms, and including 2) the subsequent identification of an appropriate measurement system and concluding with 3) the identification of specific rate and date criteria for correct and/or error performance. An example may best illustrate how each of these steps is completed in the orderly formulation of an aim standard.

Specification of Behavior in Observable Terms

Suppose an educator wanted to monitor the disruptive behaviors of a classroom of students. While most can readily imagine what a disruptive room might look like (or sound like!), without more detailed specifications, it is quite impossible to know whether the classroom disruptions result from shouting, hitting, wandering

throughout the room or lack of student response to teacher request. Not knowing a specific cause for the disruption compounds the effort to monitor effectively and therefore change the disturbing conditions. In order to monitor the behavior, therefore, (an activity which itself rests on the assumption that there is desire to change the current condition) specification(s) must be given to the term "disruptive behavior." Suppose again the teacher involved identifies "lack of student response to teacher request" as the cause of disruptive classroom activities. While that delineation does provide more information regarding the probable cause of the disruption, it is still not specific enough because it does not specify behaviors which can be observed; it does not identify behaviors as having definite beginning and ending points. The phrase "lack of student response to teacher request" can be behaviorally stated using such phrases as "Upon request, the students will sit in their seats within five seconds of the request", or more generally, "Given teacher cues or directions, students will respond appropriately within five seconds." Both statements are behaviorally correct; their difference lies in the level of specificity.

Selection of Measurement System

Once a behavior has been stated in behavioral or observable terms, the next step is to select an appropriate measurement system. Commonly used systems (or scales) of measurement include percent data, rate per minute data, duration or cumulative measures. The type of system selected is a function of the nature of skill and the desired outcome. Continuing with the example already presented, the teacher has two logical options for measurement systems: rate per day (or rate per period) and percent data. (Since he or she is not concerned that the students stay in their seats for any specific amount of time, duration was not seen as a logical option.)

In selecting one system over the other, the teacher in the example need ask only the questions, "What do I (or other teachers) desire? Will I consider that students have been successful in ameliorating disruptive behavior when they can respond correctly to 30, 40, 50 or 60 requests per day? Or is success more accurately defined in a statement of percent? Is a better measure of success the students' ability to follow requests 85, 90 or 100% of the time"? A little practicality, a little reason and a focus on the expectations of a larger society answers the question. It is, quite obviously, not important that there is a high rate of compliance; what is important is that when appropriate, the students can and do comply. In this case then, percent data is the reasonable and appropriate selection for a measurement system.

Careful selection of an appropriate and sensitive measurement system is important. While rate data are often understood to be the most sensitive type of data, in many cases, as in the example, they are not the most logical choice. Conversely, when frequency data are the logical choice because ultimate success on the skill involves fluency, then they should be used.

The proper selection of data systems should assist both teachers and students focus attention on meaningful expectations. Table 1 presents academic and social behaviors and the systems often selected to measure them. The table is not a hard and fast answer to selection dilemmas, but it does provide a point of departure for those designing systems for the first time.

Specification of Rate and Date Criteria

The specification of the level of acquisition or mastery needed to acquire or become proficient at a skill is termed a "desired rate," "desired aim" or goal. A desired date is a specification of the projected (or anticipated) date at which the student will achieve the desired rate.

TABLE 1

BEHAVIORS AND MEASUREMENT SYSTEMS

Academic Behaviors	Commonly Used Measurement Scales
Reading	
Saying words, phrases or sentences	Rate/minute
Sight Words	Rate/minute
Comprehension	Percent
Math	
Facts Practical Math	Rate/minute
Concepts	Percent, trial
Spelling	Cumulative words learned to criteria/week; percent

Social Behaviors	Commonly Used Measurement Scales
Language (appropriate interactions)	Number of appropriate and inappropriate imitations or responses/day or/play period (raw score)
On task behavior (e.g., sitting at desk)	Duration
Compliance	Percent
School Attendance	Cumulative

Desired rates. Desired rates (aka: criteria) are statements of precise standards for student peformance, such as "80% correct for two consecutive days", "100% compliance for five consecutive sessions", or "120 correct words per minute with two or fewer errors." The rates that teachers set for students (or better yet, set with students) should ultimately reflect a criterion that enables students to use or recall the learned information in an efficient manner. When students are unable to recall or use information after a specified criterion has been met, it is an indication to the instructor that the rate was not sufficient to achieve the desired behaviors. It is possible, through continued trial and error, to learn eventually which standards lead to the desired performance. But trial and error is a slow and frustrating process for both teachers and students. For this reason, a number of suggested strategies have been reported in the literature to assist students and teachers in identifying reasonable criteria. These suggestions include specifying desired rates based on: 1) levels of acceptable adult behavior; 2) satisfactory performance levels of nonhandicapped peers; 3) mean behavior scores resulting from group data; or even 4) the identification of the highest score achieved by any one peer (McGuigan, 1979; White & Haring, 1980).

There is, to date, little definitive information to suggest that any particular method of selecting desired aims is better than another. The significant factors for consideration are, however, quite clear. The established aim must result in the learner's ability to perform the desired skill independently, at a satisfactory level of competence, at the appropriate time(s) and in the appropriate setting(s). The surest evidence that a desired aim has been adequately set is the observation that the student can perform the task effectively, efficiently or both.

Desired Dates. In addition to identifying specific behavior aims, it is necessary to identify the amount of time one expects to spend on the development or

refinement of skills. The specification of this time results in a projected date for achieving the desired aim. As difficult as this is to do, establishing desired dates provides teacher and student with approximate notions of how long one can expect to take "getting from here to there." It provides the pacing element so important if yearly goals are to be achieved consequent to completing short-term objectives.

Like desired aims, establishing desired dates remains a somewhat tenuous business. Research results confirming the adequacy of a particular procedure or system are yet unknown. In absentia, however, the following formula by McGuigan (1979) may provide a starting point:

$$\frac{\text{number of teaching days}}{\text{number of teaching steps}} = \text{number of days per step}$$

An example may best illustrate the use of the formula:

> Mr. Adams knows he has 15 different competencies he wishes to teach the sophomores in the course, Basic Reading for Pleasure. Likewise, he knows, counting holidays, exams days, and so on, that he has approximately 36 teaching periods during the term. In planning the course syllabus, Mr. Adams divides 36 (the number of teaching sessions) by 15 (the number of competencies) and finds that 2.4 sessions are available per competency. He rounds 2.4 off to 2.5 periods per competency and then makes further adjustments by lessening time on less difficult or less critical competencies and adding time to more difficult or critical concepts. Through common sense calculations Mr. Adams has, in fact, established desired dates for each competency. Having established specific mastery criteria for each teaching component, he now knows not only what he wants to teach, but how long he has to do so.

121

Obviously, not every goal will be so readily amenable to such a formula. Many times the "timeline" is an already established fact. Other times, "paced time" is not even an option. If only five one-hour counseling sessions are planned, the reasonable goal for counselee and counselor must be fitted to the time frame and not vice versa. If the regular classroom curriculum mandates that 20 words must be learned per week, then that may well become the most logical starting point for setting desired dates in the support or special education classroom. The important thing to remember in establishing desired rates and dates is that both teacher and student have some notion of where one wants to go and how much time one has to get there.

Information resulting from the establishment of desired rates and calculations of desired dates makes it initially possible to pace oneself appropriately, to select, systematically, instructional strategies and materials for achieving desired rates by desired dates and to make changes when there is evidence that the current program is ineffective. Knowing precisely when a change is needed is the topic of the next section.

PROGRAM GUIDELINES AND DECISION RULES

While it is possible periodically to review collected data and subsequently make logical decisions regarding reasonable next steps, the use of program guidelines and data decision rules offers a precise alternative to random data examination. At the same time, it increases teacher efficiency. A program guideline is, most simply, a line entered on a chart indicating the projected performance path. Data decision rules are those guidelines educators may use to make decisions about when to maintain or change a particular instructional program.

Setting Program Guidelines

A program guideline (aka: performance guide, aim line) is most simply a line entered on a chart which indicates a path to be followed if a particular aim is to be achieved by a particular date. Program guidelines are most often entered on charts beginning at the level of current performance (a baseline score) and ending at the desired rate on the desired date. Figure 1 displays a program guideline as it has been entered on a chart beginning at the pretest score of 7 (a) to the desired aim of 35 (b). An "aim star" (c) is entered at the rate of 35, four weeks from the start of the program (the desired date). (The crossbar of the aim star (b) intersects the chart at the desired rate; the apex of the star (c) intersects the chart at the desired date). A program guideline (d) is then drawn from the beginning score to the desired aim.

If desired date and rate are realistically set, the program guideline provides a reasonable line of progress for expected student performance. It can be expected, of course, that performance will not fall directly on the line but rather near it. The critical question for the educator becomes, "how much deviation about the line can occur if the aim is to be met on or before the desired date"?

Decision Rules. Again, there are no tried and true research efforts to indicate that the following discussion is infallible. But work by Liberty (Note 1) and White and Haring, (1980) suggests that performance for accelerating targets which falls below the program guideline three consecutive days is not likely to alter enough to again approach or exceed the guideline. The decision rule is stated as: make a program change when data fall three consecutive days below the guideline. Conversely, if a guideline has been drawn for a decelerating target (e.g., to reduce errors or inappropriate social behaviors) a comparable rule is: change the program when data fall three days consecutively above the program guideline (see Figure 2).

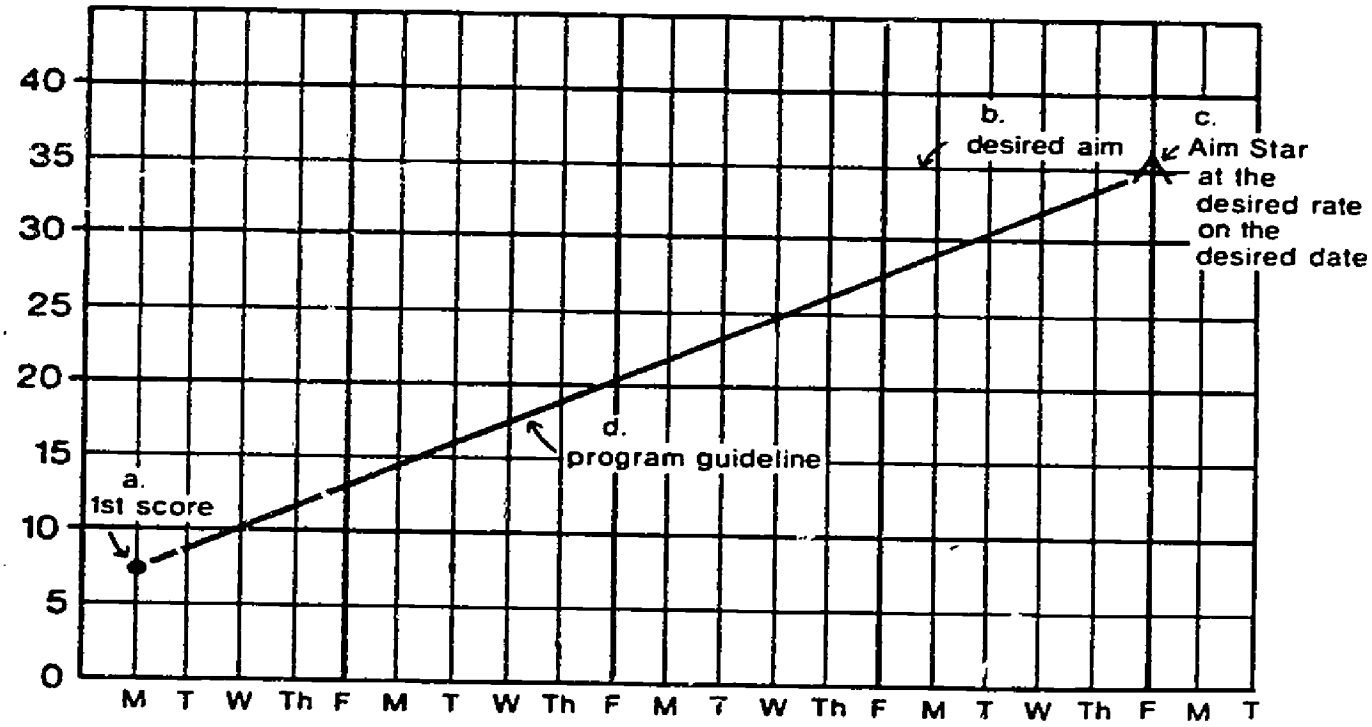


Figure 1: Program Guideline for an Accelerating Target

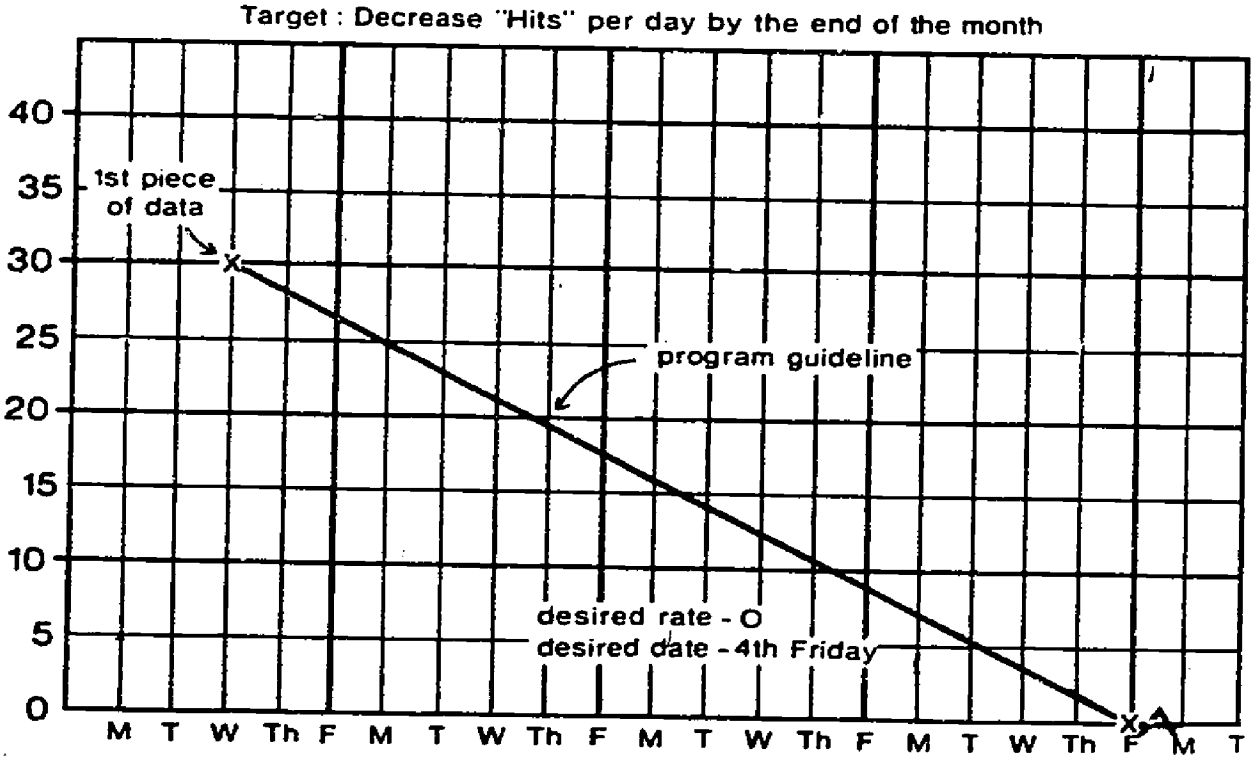


Figure 2: A Program Guideline for a Decelerating or Error Target

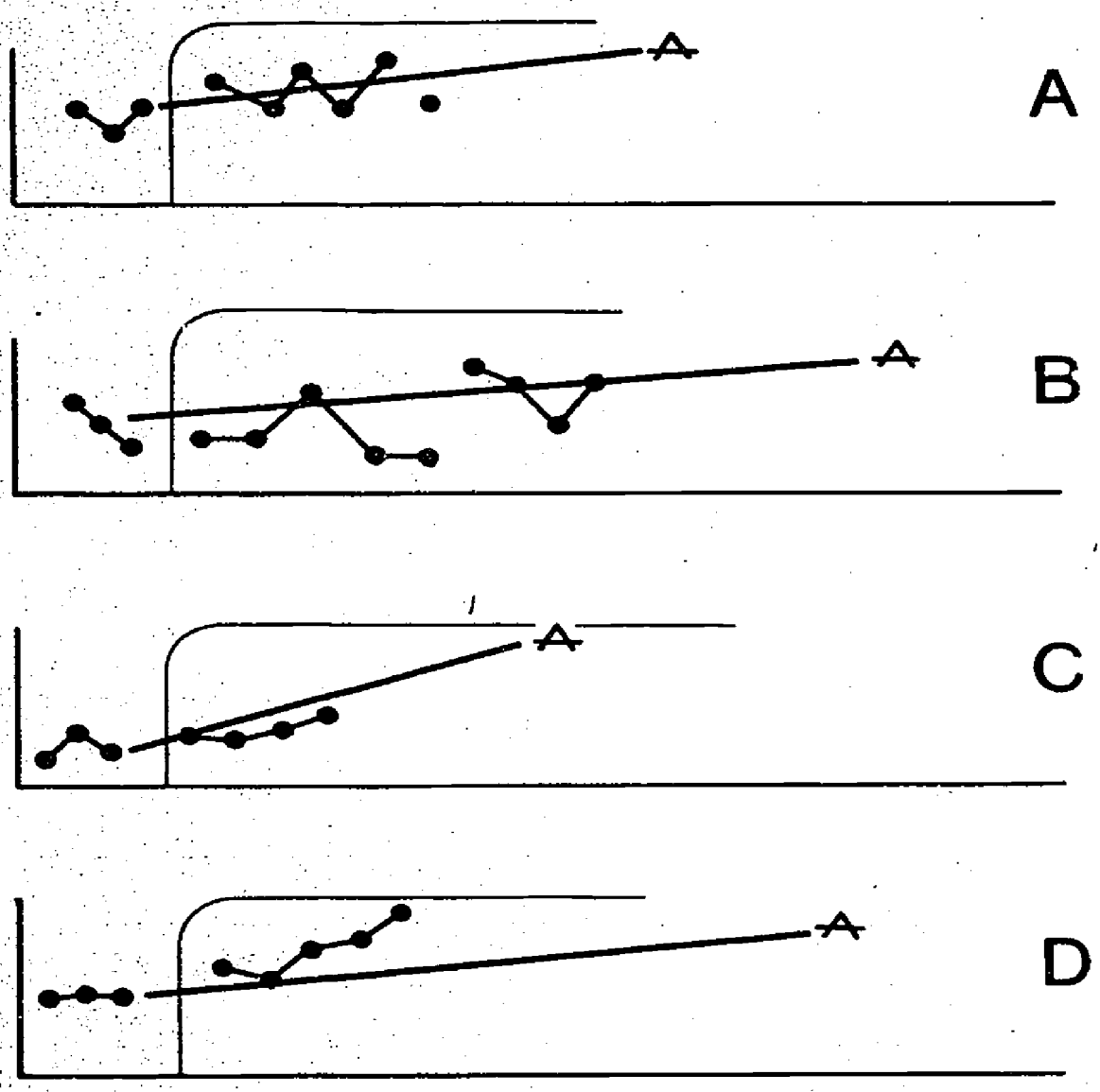


Figure 3: Sample Charts for Determining Program Change

A. No change.
B. No change.
C. Change: Data three consecutive days below guideline.
D. Change: Aim met early. Time for a new program!

The use of program guidelines for both accelerating and decelerating targets makes it possible to decide systematically when it is appropriate to continue a program or to make a program change to enhance the quality or quantity of learning. Figures 3 and 4 provide sample data sets. By studying each set, it is possible to determine on which data sets a program change is appropriate or necessary.

Examination of data sets on Figure 3 indicate that no changes are needed on data sets A and B if the three-day rule is applied. While data do bounce around the line, at no time do they fall consecutively three days below the guideline. In each case, therefore, it is appropriate for the teacher and student to continue the program as is. Applying the same three-day rule, change is called for in data set C. A program change is also appropriate in data set D, not because the data fall below the line, but because the desired aim was achieved (early).

Applying the three-consecutive-day rule for error performance on data sets in Figure 4, it again becomes apparent that changes are called for in sets B and D. In data set B, the error data fall above the guideline for three consecutive days; in data set D, the desired aim was achieved (early).

Program guidelines coupled with data decision rules indicate when a program change is needed. They do not, however, indicate what types of changes would be effective in changing inadequate performance. Knowing what kind of change to make is a result of being efficient in analyzing and interpreting data patterns -- the topic of the next section.

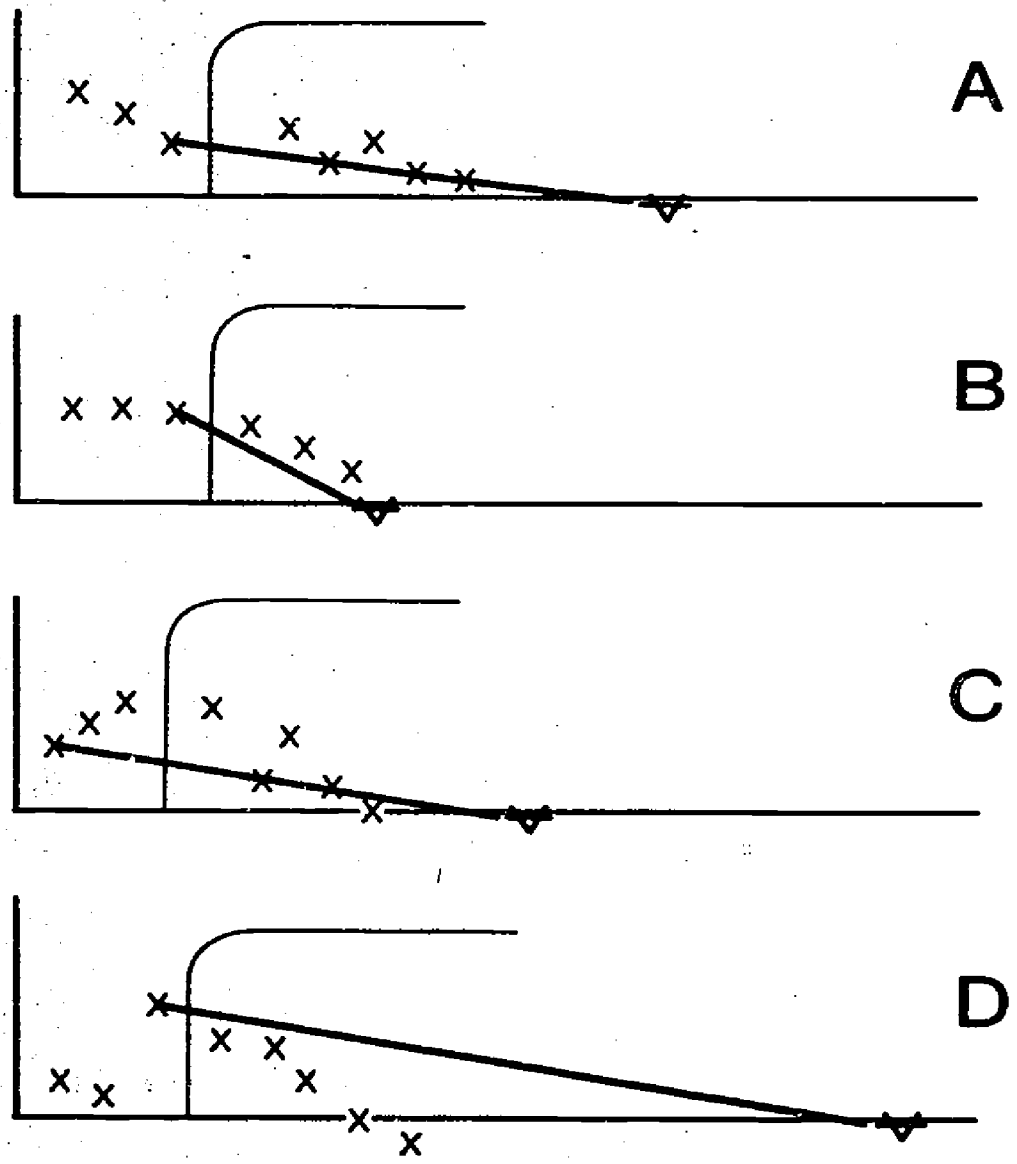


Figure 4: Sample Data for Determining Program Change for Error Responses

A. No change.
B. Change: Three-day decision rule.
C. No change.
D. Change: Error aim met.

DATA ANALYSIS PROCEDURES

It should be evident by this point that the reason for setting desired rates and dates, for employing the use of program guidelines, and for applying decision rules is to know when to make program changes. The next crucial question is what to change.

Knowing what to change is a direct function of understanding 1) the reasons for student failure and 2) data patterns. Basically, students fail to perform as desired for one of two reasons: they either are not motivated to complete the task or are not capable of completing the task given existing information. Most failures can be, upon reflection, interpreted as motivational or instructional problems. Most students do not choose to perform at a low rate; those students who do usually do not have enough information to perform faster or do not have experience with performing faster.

Failures due to lack of adequate or appropriate instruction or to lack of motivation are termed instructional problems and motivational problems respectively. Changes in programs which are made to deal with instructional problems are termed instructional interventions; those dealing with motivational problems, motivational interventions. Although it is admittedly a simplistic interpretation of often complex and subtle motivational and instructional issues, for the purpose here, these two categories will represent the two major program change categories.

Knowing when to make what type of intervention is perhaps the single most useful piece of information available to a teacher concerned with maximizing each instruction period.

Instructional Changes

The need for instructional changes is most obvious when students show a steady rate of improved performance, but a rate so low that the established aim cannot be met by the desired date. A steady, but slow or inadequate performance rate indicates that students do not have enough information to do much better (e.g., do not know the answers to certain multiplication facts, or the names of certain state capitals. Conversely, it is possible that students do have all the information necessary, but do not have a proficient grasp of the information (e.g., can figure out the answer to a multiplication problem, but must calculate it on paper; know the name of a capital, but must "think about it").

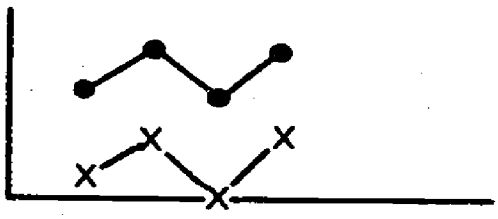
When student progress is inadequate due to lack of information, the best intervention an instructor can make is to provide the necessary information. When performance is hindered due to lack of proficiency, the most logical intervention is one composed of consistent, meaningful practice or drill. It is never to be assumed, when data patterns indicate that performance is improving -- no matter how slowly, that the problem is motivational. Other patterns are more indicative of that problem.

Other data patterns indicative of instructional problems include high correct and high error scores on the same day and low correct and low error scores on the same day. Such patterns indicate that the student can do the task accurately when working slowly, but makes errors when the pace increases. A plateauing of correct responses coupled with a gradual deceleration of errors may also be indicative of an instructional problem. With such a pattern, it can be assumed that the student is focusing attention on correcting errors. Other patterns indicative of instructional problems are presented in Figure 5. Table 2 presents a number of instructional intervention options.

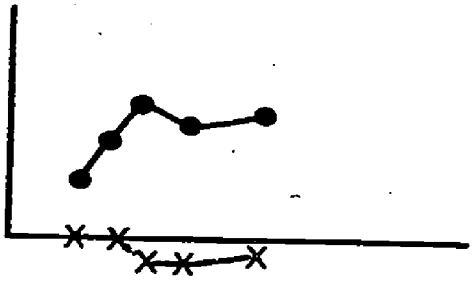
Figure 5: A Data Pattern Review Lesson

Look carefully at the data before you.

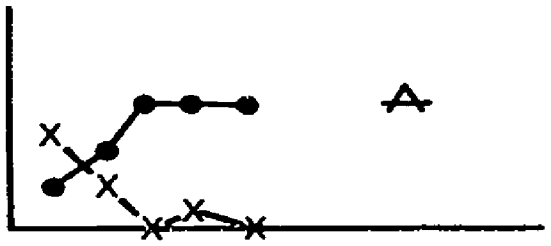
IF YOU SEE high corrects and high errors on the same days and lower errors with lower correct rates, you know the student can do the task accurately when working slowly, but makes errors when hurrying. Try an INSTRUCTIONAL CHANGE. Increase student accuracy by working for the mastery of the error items. At the same time, give the student some experience with higher rates by having him or her practice correct items.



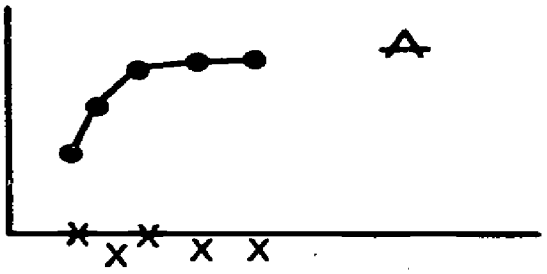
IF YOU SEE high corrects and low errors on the same day, you know the student can do the task, but sometimes chooses not to. Try a MOTIVATIONAL CHANGE.



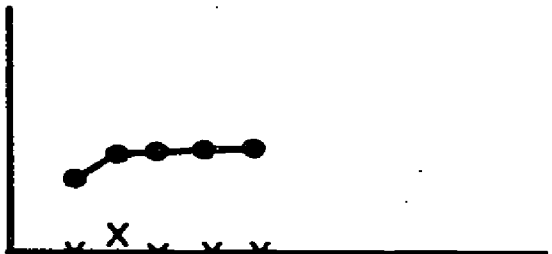
IF YOU SEE plateauing corrects and gradual deceleration of errors, maybe the student is concentrating on correcting previous errors. Continue to help the student decrease his or her error by adding specific practice on errors -- an INSTRUCTIONAL CHANGE.



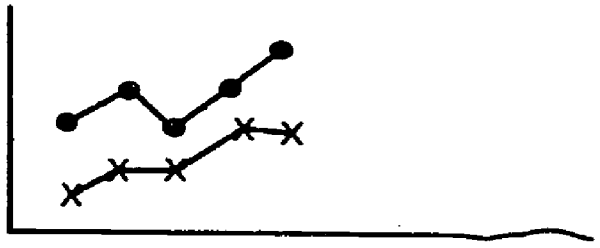
IF YOU SEE gradually increasing correct responses ending in a plateau, look at the last high rate. If it is near the desired aim, the student may be bored with the task. This is especially true if errors are at zero. Try a MOTIVATIONAL CHANGE.



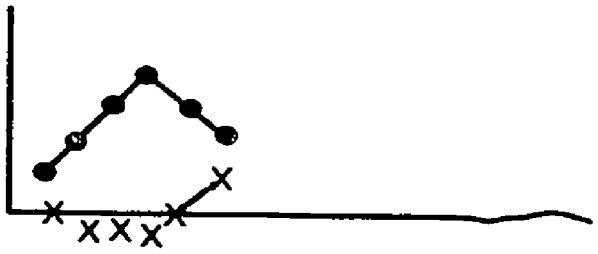
IF YOU SEE gradually increasing correct responses at a low rate of performance (not near the desired aim), check to see if the student knows anymore of the correct answers you desire, or understands that she or he should be working faster. Try an INSTRUCTIONAL CHANGE.



IF YOU SEE errors plateauing at a high rate (above 3), you know that the learner does not know the correct answer to specific items. Try an INSTRUCTIONAL CHANGE. Identify the errors and teach directly to each error item.



IF YOU SEE declining correct responses and increasing error responses, student performance is definitely getting worse. Where once there was good performance, there is now the battle between student and subject. Try a MOTIVATIONAL CHANGE. And check yourself. Have you followed through on contingencies once established?



From the Nevada Teachers' Resource Guide, Module 8, McGuigan, 1979.

TABLE 2

A DOZEN EXAMPLES OF INSTRUCTIONAL CHANGES

1. Provide more 1.1 or small group instruction
2. Increase the kind or amount of corrective feedback you are giving
3. Drill to specific errors only
4. Increase the amount of specific drill time
5. Change drill materials to the most simple, straight-forward practice sheets which have the student practice the exact response you want
6. Have the student work, drill, or practice with a peer who has already demonstrated accuracy and/or proficiency of the skill being taught and learned
7. Decrease the amount of material so that the student is working on fewer, more specific items, each session
8. Drop back to the next lowest sequence step
9. Have the student preview (read silently) the material before you begin your 1:1 instruction in reading
10. Have the student correct his or her errors three times (or five times) each before the beginning of the next activity
11. Have the student practice/drill on the basic response demanded by the skill (i.e., write numbers).
12. Have the student self-correct his or her work and self-correct (with teacher supervision) his or her own errors.

From McGuigan, 1979

Just as strategies have been found which produce increased or improved performance, a number of strategies to be avoided have likewise been identified. When making instructional changes: 1) do not change the material to "game-like activities" (the responses to the materials vary too much and student attention is drawn away from identifying the pertinent information to discovering how the task is to be completed); 2) do not set up drill sessions so that students spend the majority of time drilling on already known or mastered information; rather, focus drill work on specific errors or unknown information; and 3) do not hold back corrective or reinforcing feedback. During acquisition of new material, students need a great deal of attention.

Motivational Changes

When students demonstrate learned information, but choose not to demonstrate it consistently, it is likely that they see no reason for performing the task. When this syndrome occurs, educators need to motivate (or remotivate as the case may be) their students.

No doubt the easiest way to motivate students is to make clear the consequences of doing something or of doing something better. When changes are made to encourage students to continue improving once they have demonstrated the ability to perform the task on at least two occasions, the changes are termed motivational changes. Table 3 displays a number of optional motivational changes. Figure 5 displays data patterns indicative of a need to make motivational changes. Table 4 summarizes characteristics of both instructional and motivational problems.

These guidelines do not guarantee that students will automatically change performance in the desired direction. Many times, a number of change strategies

TABLE 3

A DOZEN EXAMPLES OF MOTIVATIONAL CHANGE

1. Decrease the amount of practice time by one minute for each additional correct response/ or decrease of one or more error responses
2. Give points or verbal praise of "beating yesterday's score"
3. Give "free time" for daily, improved performance
4. Have the student beat "the teacher's best rate"
5. Give days off (from the skill) if the skill is mastered "early" (before the anticipated date)
6. Handout "happy cards" or award citations for aims which were met
7. Keep a student chart of the skills mastered; consequate the mastery of every fifth skill by something specific the student chooses to do (but set the parameters! i.e., help the principal or the folks in the kitchen)
8. Omit practice time altogether if a student demonstrates a better rate of performance today than yesterday (they're probably practicing their heads off at home!)
9. Have the student select where he or she would like to practice and continue to let them select the location as long as performance improves
10. Have one student "race" against another student who is also trying to improve on the same or a similar skill. Whoever improves the most collects a point
11. Let the student select the person who will monitor the daily check
12. Let the student select the "best time of day" to complete work on an especially difficult task

From McGuigan, 1979

TABLE 4

CHARACTERISTICS OF STUDENTS WHO NEED:

An Instructional Change
1. Slowly increasing correct rate
2. Slowly decreasing error rate
3. Leveling-off of correct and/or error rate (scores begin to remain somewhat constant)
4. Steady increase in correct rate followed by a plateau effect or slight increase in error performance
5. Little bounce or fluctuation in the data
6. Increased correct responses but maintenance of the same error rate at a score other than one or zero

A Motivational Change
1. Correct and/or error data showing a lot of "bounce" or fluctuation
2. Random increase in error performance and/or random decrease in correct performance
3. Relatively high rate of fluency (proficiency) followed by a plateau at a rate near the criteria
4. Error at or near zero

must be tried before one is found to be effective. The preceeding guidelines simply help reduce initial programming guesswork through systematic analysis of the data and selection of probable change strategies. Ultimately, the only way to know if a particular intervention is effective for any particular student is to try it and observe the data for a period of time (usually one week) after the change is initiated. Precisely because every student is an individual and reacts individually to specific changes, the effectiveness of a strategy is always student dependent.

If the changes made for a particular student result in improved performance, it can be assumed that a correct choice was made. Many times, however, the change strategy selected will not lead to improved performance -- or not immediately. After two sessions resulting in no improved change, a different strategy may be selected. Teachers should not be discouraged if they find themselves changing programs again and again. Good teaching does not mean doing everything right the first time; it does mean having the wisdom to know when a change is needed and the competence to engage in meaningful program changes in a timely and efficient manner.

CONCLUSION

It is both possible and desirable to select specific skill criteria and to help students move efficiently toward t' criteria. It should be remembered, however, that while standards are held for students, the standards do not have to be the same for all children. (This issue alone accounts for much of the resistance to utilization of data systems in classrooms). Students are not all the same; they are innately different. The best the educational

system can do is to foster the belief that each person is unique, that difference is desirable, that no person, for possession of or for lack of skills, money, or ability to walk or to see, is any less valuable than another. The purpose of education is not to make all people the same. It is to make them better. Inasmuch as the systematic collection and use of data contribute to making individual students better, it has fulfilled its purpose, by contributing to the educational process.

The last caution is this: while there exists a very real need to engage in detailed and sensitive types of monitoring, there continues the need to adopt a more encompassing perspective. The real concern is not whether a pattern of instruction or anything else can be defined to help students gain more information. The real concern is can a pattern of instruction (or anything else) be identified which will result in students gaining the type and amount of information they need?

REFERENCE NOTE

1. Liberty, K. A. Data decision rules. Unpublished working paper, Experimental Education Unit, Child Development and Mental Retardation Center, University of Washington, 1975.

REFERENCE LIST

McGuigan, C. A. Nevada teacher's resource kit. Carson City, NV: Nevada State Department of Education, Exceptional Pupil Education, 1979.

White, O. R. and Haring N. G. Exceptional teaching (2nd ed.). Columbus: Charles E. Merrill, 1980.

Developing and Validating Assessment Instruments

Cheryl L. Hanser

A good assessment instrument is not easy to find, as anyone who has recently looked well knows. A worthwhile assessment instrument must measure what it purports to measure, accurately and reliably. It must also be suitable for the intended population. Further, it must be simple and easy to use. In order to meet these basic criteria, an assessment instrument must be subjected to an extensive process of develop - test - refine, before it is ready for publication and ultimate distribution. This process can take years - more years than are available to the average federally funded project. For most projects, assessment is the first step toward program implementation. Without initial assessment, the major thrust of the program, intervention, cannot proceed. Assessment is also crucial for program validation efforts. Without accurate, viable assessment measures of child change, the project has little hope of making a major impact on the field of education. For these reasons assessment is not only of critical concern to a project, it is also of immediate concern -- critical in that without assessment, validation of programmatic efforts is impaired -- immediate in that most projects have only three years to complete their task.

Considering the difficulties of developing and validating worthwhile assessment instruments, why then do they continue to proliferate? Simply stated, most of us get into the business of developing new assessment instruments because the existing ones do not adequately meet our needs. Either they don't measure the skills that we are interested in, and/or they lack the sensitivity to measure changes in performance.

Many assessment instruments are deficient because they don't measure the right skills. Locating assessment instruments which test what is being taught is a primary problem. Perhaps the test developers did a poor job of construct validity and did not include salient variables which define a particular skill area. Construct validity problems occur when developers fail to survey adequately a skill area prior to designing a test. For example, a test of community mobility skills would be incomplete without a section on bus riding behaviors. Construct validity problems may also occur because the test was developed for a different population. A common difficulty experienced in this regard is when tests developed for normal or mildly handicapped youth are used with severely handicapped persons.

Another major problem with many assessment instruments is their lack of sensitivity to performance changes. Insensitive tests increase the difficulty of showing effects of a particular intervention. If a project cannot measure the impact of an intervention, it will have difficulty convincing others of its merits. Sensitivity is a function of the breadth of a test and the type of response required. If a test covers a large number of objectives, as is frequently the case with achievement tests, changes in level of performance must be large in order to affect the results. Similarly, if the required response is subject to a large degree of fluctuation (i.e., is not reliable), as is the case with poorly defined items in many behavioral checklists, then changes in performance may be a function of differences in interpretation rather than differences in performance.

Given that these problems occur, what options are available? Essentially, three courses of action are possible. First, principles can be compromised and a less than perfect assessment instrument used. This option is viable when the consequences for making incorrect judgments are not severe or when small discrepancies exist between the real and ideal. For example, if the purpose of an assessment is to identify which phonic elements a student needs to learn, misidentification of a few phonic elements would not severely disrupt the intervention because the teacher could easily adapt instructional strategies. Conversely, if the purpose of an assessment is to select a student for special education, the consequences of misdiagnosis may be severe and long-lasting.

A second alternative is to modify or adapt the device. Some assessment devices can be divided into subsections and administered separately without violating reliability. Similarly, some assessment devices can be modified to facilitate responding without seriously weakening their predictive powers. Examples of modifying administrative procedures include allowing verbal rather than written responses or ignoring the allowable time limits. In any case, if an assessment device is modified or adapted, the results must be cautiously interpreted.

A third alternative is to develop a new assessment instrument. This alternative is the subject of the present chapter. It is suggested as a last resort, to be employed only after the previous options have been found insufficient.

Assessment Considerations

Five variables should be taken into account when devising a new assessment instrument. Careful consideration of these variables will contribute to the development of an

appropriate instrument. These five variables are the purpose of the assessment, what to assess, who will assess, how the assessment will be conducted, and whom the assessment results are for.

Purpose of assessment. The type of assessment chosen and its breadth will depend to a large degree on the purpose of the assessment and how the results will be used. Cone and Hawkins (1977) suggest five phases of assessment to be considered. These are screening, placement, intervening, monitoring progress, and following-up. The first phase involves screening and disposition. In this phase general types of problems are identified and areas of difficulty determined. Assessment methods are broad-band and often of low fidelity. Common methods used for screening are interviews, wide-ranging problem checklists and achievement or intelligence tests.

The second phase of an assessment, placement, is intended to place a child in a classroom or within an instructional group. In many instances, screening and placement are considered simultaneously. Individual Education Plans are usually devised during this phase. Common techniques include standardized tests.

In the third phase, specific behaviors are targeted and appropriate interventions planned. In essence, this is when baseline measures are obtained. Thus, self-monitoring, behavior checklists, analogue assessments and assessments in the natural environment may be indicated.

The fourth phase occurs when progress toward skills is monitored. At this level of assessment, measures need to be continuous and economical. Further, these measures should be sufficiently sensitive to illustrate that change is occuring.

The fifth and final assessment phase pertains to obtaining a follow-up measure of performance. It is not sufficient to change a behavior or to teach a new behavior; the

ultimate goal is for that behavior to become ingrained into a person's repertoire and, where appropriate, to be generalized into other situations. Thus, measures of performance should be obtained after a treatment has been discontinued.

Hopefully, every model project will address each phase of assessment within its design. When all phases are included, assessment is viewed as a continuum of effort rather than a periodic activity. Further, when all five phases of assessment are considered, it is readily apparent that just one assessment instrument or method is insufficient to meet all of these needs. Therefore, to improve an assessment plan, the first step is to analyze the current assessment plan and identify which assessment steps need to be developed in order to result in a complete package. Once "holes" in the assessment package are identified they can be systematically reduced through adding more assessment devices or through developing new ones. This analysis might also reveal areas in which too many assessment devises or inappropriate devices are used.

What to assess. This is the core of any assessment instrument. The what refers to the area of interest (e.g., self-help) and the specific items to be assessed (e.g., dressing). If the items adequately represent the area of interest, the test is said to have achieved content validity. Deciding what to assess is usually not a major problem for educators. Unfortunately, it is not as easy as it appears. The major content problems are the result of superficiality or incompleteness. Superficiality occurs when a test developer attempts to include too many objectives within one test. If too many objectives are covered, assessment of each objective will, of necessity, be minimal. Thus, an assessment device designed to monitor progress in self-help skills, would only obtain measures of those specific skills currently being taught, not the entire range of self-help skills. Tests may also suffer from incompleteness in that critical objectives or skills within a sequence are omitted. One example of an

incomplete assessment device would be a language test which measured only expressive language skills. Incompleteness occurs when a test is not carefully conceived or when only one viewpoint is obtained. To summarize, when thinking about the "what" of assessment, consider the skill categories to be assessed and the number of samples per category to be included.

When deciding what to assess, it is a good idea to survey similar assessment devices and curricula to determine which tasks others consider important. Another idea would be to share the initial outline with a variety of other persons and obtain their critical input. Fremer (1974) states three purposes for having test items reviewed: 1) certification that the items are appropriate measures of the objectives of interest, 2) assurance of consistency of style and clarity of expression, and 3) acceptance of the purposes and procedures by the intended audience (i.e., parents and students, if appropriate). It is far better to take the time initially to assure content validity than to wait until after the assessment device is completed.

Who to assess. The characteristics of the population for which the assessment device is intended must be considered. Some crucial variables include age, functioning level and special psychological or physiological constraints. Obviously, if the test is to be used with very young children, then simple, concrete objects or pictorial representations are preferred over written stimulus materials. Conversely, teen-aged, mildly retarded youth should not be subjected to assessment devices obviously developed for a much younger population. Alternatively, if the test is to be used with the visually handicapped, other modifications will be necessary. It is wise to remember, however, that the target population of the test should be sufficiently restricted to meet a projects' needs but sufficiently broad to be applicable to other populations.

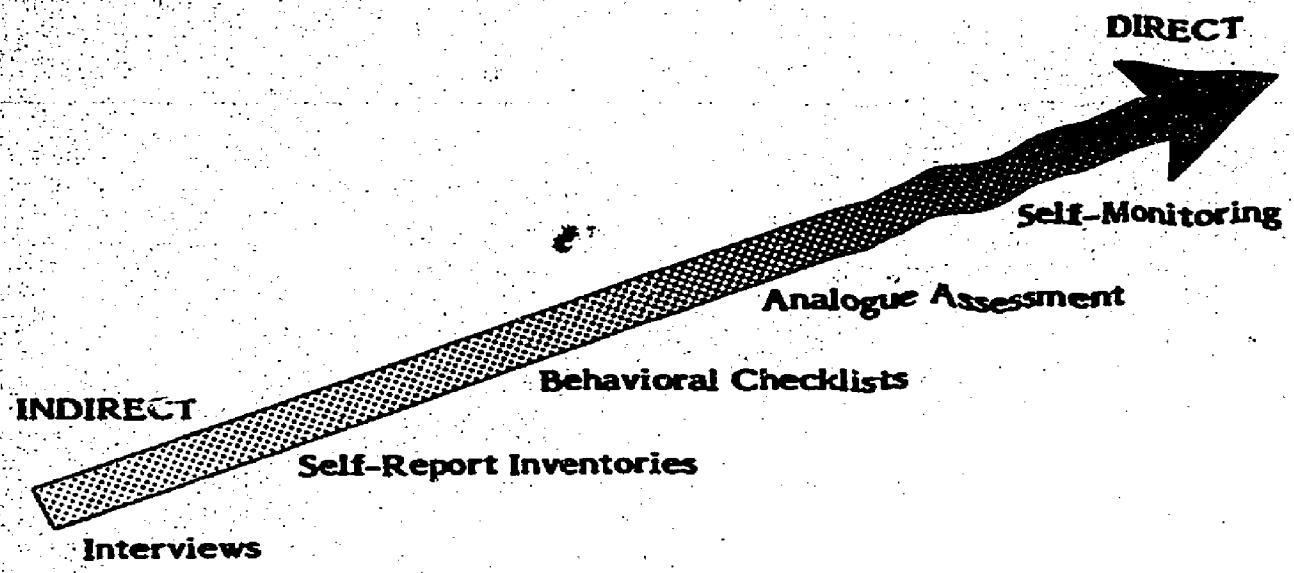


Figure 1: **A Continuum of Measurement Techniques Which Reflect the Most Indirect to the Most Direct Measures of Skills and Behaviors as They Occur in the Natural Environment.**

How to assess. This question actually involves a number of issues such as the format of the assessment instrument, directions for administration and the reliability of the measure. Each of these questions should be addressed.

The format of the assessment instrument is a primary consideration because behaviors can be measured in a variety of ways. The most direct and valid technique would be to observe the student within the natural environment. Unfortunately, due to limitations of time and resources, direct observations are not always possible. Measurement techniques can be arranged on a continuum of directness (Figure 1). It is the task of the developer to choose the technique which appears to be most direct, one which is appropriate to the behavior of concern.

The least direct measurement technique is an interview. Asking a person questions about his skills or asking how he would behave in certain situations is notoriously unreliable (it may also be impossible if the person has limited communication skills). The person may forget, or remember selectively. The person may not understand the question or may not wish to respond. Finally, the subjectivity of behavioral interviews contributes to their unreliability. It is for these reasons that interviews are seldom used in isolation.

Self-report inventories include any direct, written or verbal voluntary presentations by the subject. Thus, criterion-referenced, multiple choice, and normative tests are considered variations of self-reports. The validity of self-report data is based on the degree to which the reports correspond with actual responses (Bellack & Hersen, 1977). It is also a function of the validity and reliability of the testing instrument.

A common type of self-report instrument is the criterion-referenced test. Criterion-referenced tests yield information about the competence of individuals relative to specified instructional performance tasks. Developers are encouraged to follow a five-step process in developing criterion-referenced tests (Hambleton et al., 1975). These five steps are task analysis, definition of the content domain, generation of referenced items, item analysis and item selection. If these steps are carefully followed, content validity will be enhanced.

Behavioral checklists represent one of the most popular methods for assessing skill levels. They serve two primary purposes: description of an individuals' current skill repertoire and prescription of remediation strategies (Walls, Werner & Bacon, 1976). Checklists come in a variety of forms from simple yes/no questionnaires to complex instruments with Likert scales. Behavioral checklists can suffer from a lack of content validity if the items are not representative. In order to improve content validity and to achieve adequate response definitions, the

items should have at least three somewhat overlapping characteristics: objectivity, clarity and completeness (Hawkins and Dobes, 1977). In order to be objective, the items must refer to observable characteristics in the student. For clarity, an item must be readable and unambiguous. Finally, a complete item delineates the boundaries of inclusion and exclusion.

Analogue assessments include observations of behaviors obtained in controlled environments. Role playing and simulations are prime examples of analogue assessments. Analogue assessments are extremely useful for observing social skills. They may also be helpful when conducting follow-up assessments of adaptive living skills. It should be remembered, however, that these techniques are valid insofar as the results are representataive of actual behaviors in natural settings. Thus, whenever possible, unobtrusive measures should be obtained in preference to more obtrusive measures. Other critical concerns pertaining to analogue assessments include developing precise, objective behavioral definitions, and adequately training reliable observors.

Self-monitoring occurs when the subject counts and reports his own data. It can be a direct measure if the subject reports data truthfully and if the behavior being measured is objectively defined. Thus, the number of bites of food one consumes a day is a more objective measure than a person's food cravings. The number of bites actually consumed compared to the number reported is, of course, a issue of veracity.

The previous assessment techniques represent the more frequently used methods for obtaining behavioral data. Hopefully, test developers will choose the most direct technique appropriate to the task. Once the type of assessment technique is chosen, the next questions pertain to administration issues.

Administrative considerations include: who will administer the test, how clear the instructions are, and

how much time is available for assessment and analysis activities. If a trained, experienced clinician will be administering the test, he can be expected to make inferences based on the quality and quantity of responses. Flexible subjective assessment techniques can be utilized. Alternately, if a teacher is the intended test administrator, then the technique must take into account the constraints upon his time and the fact that he probably is less experienced than a clinician. Finally, if a paraprofessional and/or parent will be giving the test, then it should be quick and simple.

Test directions must be carefully described in any assessment device. Directions should be written clearly so they can be easily understood. They should describe the purpose of the test, the situational variables to be controlled, specific directions for delivering the salient stimuli and recording the response, definitions of correct and incorrect responses, procedures for reinforcement (if appropriate), and procedures for corrections. If possible, the assessors should be trained prior to the actual test administration to ensure the reliability of the testing procedures. This prior training is especially crucial when the assessors are required to use behavioral observation techniques.

The time available for administration and analysis of the test results is an important consideration. A long test is more difficult to schedule than a short one. The length of a test also affects the reliability of the data because the subject can become fatigued. Further, if complex computer calculations are required, then the time between initial data collection and completion of the final results will increase.

The reliability of a response is a function of the method in which the item was written, the response measured and the assessor trained. A poorly written item is unreliable because it is open to different interpretations. An infrequently or inadequately measured response is unreliable because it may not reflect the student's actual

performance level. Finally, poor training affects reliability because assessors may not accurately observe or record all occurrences of a behavior.

Whom the assessment results are for. The final question to be addressed pertains to the reporting procedures. The results of an assessment may serve many purposes. They may be used to plan and implement a program of instruction. They may also be used to document skill improvements. Or else they may be used to convince others of a program's value.

The report format will vary depending on the audience for which the results are intended. Funding sources, such as legislators and the Bureau of Education for the Handicapped, are interested in summative data presented in charts and tables. Parents and teachers, on the other hand, may also be interested in summative data, but on an individual basis. The presentation of the data should also be modified in accordance with the background of the receiving party. Professionals reading results in journals are not offended by jargon. Parents are not only offended, they are often very confused. In any case, few people are interested in all of the available data; therefore, be selective. Provide the least amount of data which tells the most pertinent facts about a program. Whenever possible, graphs and summary tables should be used.

These five variables should be considered when devising an assessment instrument. They are not steps in the usual sense; rather, they are concerns which interrelate and impinge on each other. Some concerns, such as the characteristics of the population, are essential. Others, such as properly training the assessors, are important because they increase the reliability of the measures.

Test developers need to be aware that many compromises occur in devising an instrument. For example, a test developer may decide that ten observations should be obtained of a student's social interaction skills over a

month's period of time. Further, he may decide to observe the student in a variety of natural settings. These data would provide strong evidence regarding level of performance on a variety of social indices. However, if time is minimal or staff negligible, a more realistic assessment strategy might be to obtain three observations in different analogue situations combined with a structured interview. While the resultant data might not be as convincing, at least they will be available.

Test developers should also remember that devising assessment instruments consumes a considerable amount of time and effort. The results of one field test are used to modify the instrument so that it can be field tested again and again modified. Oftentimes it seems as if the assessment device is forever in a developmental stage. Perhaps that is inevitable or even desirable. The following example illustrates the process by which one assessment instrument was developed and refined.

AN EXAMPLE

A number of years ago, a series of curriculum research studies was conducted with intermediate-aged learning disabled students at the Experimental Education Unit, University of Washington, Seattle. Prior to conducting the studies, it was necessary to place the students appropriately in reading, math and spelling materials. The Reading Placement Inventory (Lovitt and Hansen, 1976) was one result of these placement efforts.

The original intent was not to develop an assessment instrument. The original intent was to locate and use a commercially available assessment instrument which directly measured contextual reading skills, and which reliably predicted a student's performance in a given

reader. The first step, then, was to review the literature to locate appropriate assessment instruments. Four common methods used for reading placement were identified. These methods were to: 1) assign readers on the basis of chronological age; 2) assign texts on the basis of achievement test scores; 3) assign texts on the basis of a reading placement test which accompanies a series; and 4) assign texts based on the results of an Informal Reading Inventory (IRI) score. Of the four methods, the Informal Reading Inventory was the most direct; however, because only one sample per reading level was obtained (hence reducing its reliability), the decision was made to modify this technique.

Construct validity was assured with the IRI because students were assessed and instructed in the same materials. For the Reading Placement Inventory, the reliability of the results was increased by obtaining more than one reading sample per level. Reading samples were obtained from the beginning, middle and end of each textbook, a total of five samples per level. Other modifications of the IRI procedure included standardizing the length of reading samples and the number of comprehension questions. Finally, the decision rules used to place students instructionally were changed.

The question of "who" to assess was already determined when the students entered the program. The students, ranging in age from 8 to 13 years, were of normal intelligence but were performing academically between one to three years below grade level. None of the students exhibited obvious sensory or behavioral impairments; however, all students had at least some minimal reading skills. Thus, an assessment of contextual reading skills was appropriate for this group.

In considering the "hows" of assessment, the format chosen was a type of self-report inventory. The students read the passages and responded orally to the comprehension questions. The assessment was administered by the classroom teacher who had three

years' experience teaching reading. To increase the reliability of the measures, a second person obtained reliability measures on the procedures, responses and data recording according to a predetermined schedule. Finally, all of the directions used for administration, recording responses and analyzing the results were available in written form.

The assessment results were used in a variety of ways. Initially, the results were used to assign textbooks. The assessment was readministered at the end of the school year and the results of both administrations used to determine the magnitude of the student's progress during the year. The assessment results were also compared with the students' subsequent performance in their readers to verify the stability and predictability of the assessment device.

The Reading Placement Inventory was found to be a useful assessment instrument. With it, students were placed in readers with a great deal of accuracy. The data were also used as part of a summative evaluation. Before the inventory itself was ready to be shared, however, it was subjected to numerous revisions.

One concern was in regard to the ability of the Reading Placement Inventory to predict a student's reading level. After placement, only one student in two years had to be reassigned. The remaining 13 students appeared to be correctly placed. Further, after placement all students progressed at satisfactory rates. Thus, it appeared that the type of assessment and the decision rules used were satisfactory. The predictability of the reading rates was also of interest. The students' reading rates (correct and incorrect) were compared during placement and in the first week of instruction. These rates were similar. However, when the percentage of correctly answered comprehension questions was compared for these two intervals, a decrease of nearly 15% was observed. This problem was subsequently resolved by requiring the response mode (either oral or silent) to be consistent for

placement and instruction. A final concern was the amount of administration time necessary. The original procedure was reliable but took about eight hours per student to complete. This was felt to be too time consuming for the average teacher. One way to reduce this burden was to reduce the number of reading samples obtained. Therefore, the predictability of the assessment was compared for different numbers of samples per readers. When the data were analyzed in this manner, three reading samples per grade level were found to be as reliable as five samples. One reading sample per grade level, however, was an unreliable predictor of reading level.

Based on these findings, the assessment instrument was modified and disseminated. First, the number of reading samples per book was reduced from five to three. Second, the mode of response for the comprehension questions was changed so that the students read the questions and responded to them in writing. Finally, it was determined that the students needed to read only a sample of readers, not all of the books.

CONCLUSION

Developing an assessment instrument is not an activity to be taken lightly. A good assessment instrument is the product of careful development, field testing and refinement. It is valid, reliable, useful and desirable. It is valid only to the extent that it accurately measures what it purports to measure. It is reliable only to the extent that the behavior measured truly reflects reality. It is useful only to the extent that it is sensitive to changes in performance. Finally, it is desirable only to the extent that it measures behaviors which society cherishes.

Educators will continue to develop assessment instruments and to modify existing ones. Unfortunately, many of these devices will be as unsuitable as the devices they seek to replace. Quality control is required for new assessment instruments. Anyone who develops an assessment instrument should apply the same criterion to his instrument, as he applies to commercially developed tests. When developing a test, the purpose of the assessment, what will be assessed, who will be assessed, how the assessment will be conducted and whom the assessment results are for must be considered. However, test development should not stop there. Any new assessment device should be carefully field tested and refined before it is shared with others. Only by carefully monitoring ourselves can we hope to stem the flow of invalid and unreliable assessment instruments.

REFERENCE LIST

Bellack, A.S. & Hersen, M. Self-report inventories in behavioral assessment. In J. D. Cone & Hawkins R. P. Behavioral assessment: New directions in clinical psychology. New York: Bruner/Mazel, 1977.

Cone, J.D. & Hawkins, R.P. Behavioral assessment: New directions in clinical psychology. New York: Bruner/Mazel, 1977.

Fremer, J. Developing tests for assessment programs: Issues and suggested procedures. Princeton, N.J.: Educational Testing Service, 1974. (ERIC Document Reproduction Service No. ED 093-990).

Hambleton, R.K., Swaminathan, H., Algina, J. & Coulson, D. Criterion referenced testing & measurement: A review of technical issues and developments. Paper presented at the meeting of the American Educational Research Association, Washington, D.C., 1975.

Hawkins, R.P. & Dobes, R.W. Behavioral definitions in applied behavior analysis. Explicit or implicit. In B.C. Etzel, J.M. LeBlanc & D. M. Baer (Eds.), New developments in behavioral research: Theory, method and application. In honor of Sidney W. Bijon. Hillsdale, N.J.: Lawrence Elbaum Association, 1977.

Lovitt, T.C. & Hansen, C.L., Round one - placing the child in the right reader. Journal of Learning Disabilities, 1976, 9, 18-24.

Walls, R.T., Werner, T.J. & Bacon, A. Behavior Checklists. Morgantown, WV: West Virginia Research and Training Center, 1976.

Concluding Remarks: Using Assessment Data to Document Program Effectiveness

Cheryl L. Hansen

The assessment process supplies the data which become the basis for many educational decisions. Assessment information is used to plan, implement and refine programs for children and youth. This use of assessment information constitutes a valuable and necessary activity. Model programs, however, have an additional goal. Namely, they are charged with demonstrating, validating and replicating exemplary educational systems for handicapped children and youth. Thus, model programs must not only obtain data with which to make programmatic decisions, but these data must be obtained in such a manner as to convince others that the program was responsible for the learning which occurred.

Assessment data provide the foundation of program validation efforts. These data are a necessary prerequisite for convincing others of a program's merit; however, they must be supported by data from other sources. Supporting data might include information on costs, numbers of students served, degree of support obtained from parents and so forth. Supporting data might also be obtained to compare the target students' performance with that of another group of students. All

these types of data are used to document program effectiveness. Collectively, they are known as program evaluation activities.

The precise nature of data obtained for program validation depends, to a large extent, on the intended audience. Thus, in order to be convincing, one must identify the potential constituents one wishes to address and then identify the type of information that will be most convincing. Prior to identifying potential constituents, a distinction should be made between assessment and evaluation activities. Assessment and evaluation share three facets: rationale, measurement and judgment (Adelman & Taylor, 1979). Although they share these facets, each is operationalized in a different manner.

In both instances, there is a rationale, a reason for the activity. In assessment, the rationale might be to identify the general nature of service needed, to develop a specific intervention plan or to make a diagnostic classification. Alternately, in evaluation, the reason is often to specify the impact of a particular intervention and to rule out competing hypotheses for the observed results.

Assessment and evaluation also share common measurement functions. In both instances the common denominator is the data which are obtained. The adequacy of the measure is critical to the subsequent judgment to be made. Both assessment and evaluation data are obtained on the actual skills or behaviors exhibited by a student. The difference between assessment and evaluation measures, therefore, is the supporting information. For assessment, the student's behavior -- its frequency duration and/or intensity -- is sufficient for making decisions. For evaluation, however, these data must be compared with other measures or with data from other students. Thus, the student's behavior must be compared with that of a peer, a normative group, or with his own baseline behavior to assess the

significance of the behavior change. Evaluation activities might also necessitate obtaining other indices of behavior change. For example, school records, questionnaires and self-report inventories may provide information about the impact of a particular program. These data would not provide diagnostic information (i.e., what skills a student possessess). They would, however, provide environmental impact data and thus could be presented as one part of an evaluation packet. Therefore, while assessment and evaluation activities begin with the same data base, evaluation activities often proceed a step further to include additional data.

Finally, assessment and evaluation are used for decision making. The judgments which result from assessment and evaluation activities reflect the initial rationales. Assessment data might be used to initiate a service delivery plan, to provide specific instruction on a skill, or to choose between intervention techniques. Each of these decisions affects the educational program of an individual student. The results of evaluation activities, on the other hand, might be used to continue a total service program in a school district, to provide additional support for such a program, or to choose between competing methodologies. The impact of these decisions is also different. Assessment decisions are child based. They affect the life of one child within a particular program or set of program options. The impact of evaluation decisions, by contrast, may affect a group of children, or it may affect the widespread dissemination and adoption of a particular methodology.

Thus, it is incorrect to use the terms assessment and evaluation interchangeably. They are mutually dependent but not mutually exclusive. Assessment can occur in the absence of evaluation. This is exemplified by the teacher administering an Informal Reading Inventory. However, evaluation cannot occur without assessment. To complete the example, the results of the Informal Reading Inventory cannot be interpreted to suggest that the teacher's instruction was solely responsible for changes in the child's performance.

Child assessment is the core, the foundation of a program, but it provides only one bit of evidence for proving program effectiveness. In order to determine which other types of data are necessary for evaluating a program's worth, it is necessary to look closely at the characteristics of the constituency and to determine what questions need to be answered. Once the questions are identified, the search for measures and experimental designs can ensue.

IDENTIFYING THE QUESTION

Evaluation data are requested by many audiences which have unique questions and concerns. These audiences are students, teachers, program developers, parents, other educators, program evaluators, legislators and philosophers. By carefully identifying these eight audiences, delineating their concerns, and obtaining relevant data, the project director can be assured of establishing a well-rounded evaluation plan. Table 1 lists the constituents, their concerns, and potential sources of data used to respond to those concerns.

The basic response of the student, when faced with the successful accomplishment of a newly acquired skill, is often one of relief. The student is only interested in whether he learned the skill. He is relieved that it will no longer be necessary for him to receive instruction -- no longer will he be the only kid in his class who can't (fill in the blank). Along with relief, the student might feel pride that he accomplished the skill, or curiosity regarding the next skill to be tackled. The primary sources of data used to answer the student's concerns are direct, repeated measures of his performance on the instructional task. These data are usually compared to some mastery criterion.

TABLE 1

CONSTITUENTS, CONCERNS AND DATA

BASES USED TO DOCUMENT PROGRAM EFFECTIVENESS

Constituent	Concerns	Data Base
Student	Did I learn?	1. Direct, repeated measures/criterion-referenced/accuracy
Teacher	How well did he learn? How does he learn best? What skill is next?	1. Accuracy/fluency on direct, repeated measures 2. Diagnostic analysis of learning environment 3. Refer to curriculum/content analysis of skill/IEP
Program Developer	How well did he learn? Is there a more efficient instructional method?	1. Accuracy/fluency 2. Item analysis 3. Diagnostic analysis of learning environment 4. Time/trials to criterion 5. Probes for skill generalization and maintenance
Parent	How well did he learn? What is his potential?	1. Accuracy/fluency 2. Learning rate/comparisons
Other Educators	How well did he learn? Is "he" similar to my students?	1. Accuracy/fluency 2. Demographic/gross performance measures 3. Cost of materials, staff training required, physical space needed, time and effort.
Program Evaluator	How well did he learn? How valid and reliable were the measures? To what can the learning be attributed?	1. Accuracy/fluency 2. Analysis of assessment instrument and administration procedures 3. Analysis of threats to validity
Legislator	How well did he learn? Is the program cost efficient? Is the program worthwhile?	1. Accuracy/fluency 2. Cost of implementation in time and resources as compared to number of pupils served. 3. Impact measures such as ratings of consumer satisfaction, documentation of interest generated (e.g., number of visitors, number of requests for information), successful program replications completed.
Philosopher	Who cares? Are the goals appropriate? Do the ends justify the means?	1. Ratings of significant/appropriate experts 2. Numbers of replications begun/continued 3. Ratings of satisfaction by consumers, parents, students

The student's teacher might also express a degree of relief. He is relieved to know that the skill has been mastered and is looking forward to teaching a new skill. The teacher also wants assurance that the student can perform the skill at a proficient level, one which will permit him to use the skill to learn other skills. Further, the teacher is concerned that the skill is sufficiently ingrained that the student will not forget it, thus necessitating future reteaching. Finally, the teacher is always analyzing and systematically modifying the instructional environment to facilitate optimum learning

conditions. Thus, the teacher wants to know the accuracy and fluency with which the student preforms a task, the conditions which facilitate skill acquisition and maintenance and what task should be taught next.

The program developer shares some concerns with the student and teacher. He is interested in the fact that the skill has been acquired. He is also interested in the proficiency and retention of the skill over time. In addition to these concerns, however, the program developer is extremely interested in whether the skill could have been taught more efficiently. He is interested in the student's responses to each frame of instruction and the student's rate of acquisition. In other words, the program developer needs to analyze the student's daily performance and to modify systematically the instructional materials in order to teach the skill more quickly, or with fewer errors, or to a more advanced level. Finally, the program developer uses instructional probes throughout instruction to test for skill generalization and maintenance.

The parents are, of course, interested spectators (or in some cases participants) in the learning process. Their concerns are many. They want to know whether the skill was learned and whether it will generalize to other environments (i.e., will he talk that way at home). They may wish to make comparisons between the student's performance and that of his siblings or peers. Most importantly, the parents are interested in prediction. They often want to know how much the child will be capable of learning. Similarly, they may be interested in determining how long the child will need to participate in special education programs. Answers to these questions are seldom simple. They require comparing an individual's learning rate and performance with another standard of criterion and making a "best guess".

Another educator is primarily interested in the impact of the total program, rather than the performance of an individual child. He wants to be convinced that the

program teaches what it says it will teach. He also wants documentation to show that it can work with his students (i.e., demographic information). Further, another teacher needs to know how to implement the program in his own classroom. In this regard, other teachers and administrators require information pertaining to costs of materials, staff training required, physical space needed, and amount of time and effort expended. Other teachers do not necessarily need to be convinced that the program is the best possible, only that it is better than their current program.

The program evaluator is a pessimist. He does not believe that the program was responsible for the observed performance changes. The program evaluator wants to determine whether or not these students learned because they just got older, the teaching was novel, the test was administered in a biased fashion or because of some other factor. Thus, the program evaluator attempts to discern a functional relationship between the learning and the instruction. He is concerned with the validity and reliability of the assessment techniques and with the adequacy of the experimental design.

The legislator has simple, but crucial, concerns. The legislator merely wants to know whether the program was worth the monetary investment; his concerns are of crucial importance because he holds the purse strings. The legislator wants a gross measure of program impact. He wants to know: 1) Did the program teach the students anything, 2) Did the students acquire the information in a cost-efficient manner, and 3) Is the program worthwhile? The types of information of interest to the legislator (in addition to accuracy and fluency data) are the number of students served, teachers trained or skills learned. In addition, legislators may be interested in impact measures such as the number of visitors at a site or the number of requests received for information. Finally, legislators are often convinced by testimonials of program effectiveness given by educators, parents or other "respectable" authorities.

The philosopher (and there is a little in all of us) asks the enduring questions. He has three basic concerns: 1) Are the goals appropriate?, 2) Do the ends justify the means?, and 3) Are the consumers satisfied? (Wolf, 1978). These questions relate to the social validation of the program. They can be resolved by asking educators, parents or other authorities to rate a program's appropriateness and by asking consumers to rate their satisfaction. Other sources of data could include measures of the interest generated by a project and numbers of replications initiated at other educational sites.

These eight audiences: student, teacher, program developer, parent, another educator, program evaluator, legislator and philosopher, are each interested in making decisions about our programs. Their decisions may affect the quality -- possibly the provision -- of educational services for handicapped children and youth. It is imperative, therefore, that the needs of these different audiences are correctly matched with the most accurate, relevant data possible.

CHOOSING A MEASURE

A good measure is persuasive. It supplies evidence in sufficient detail to persuade others of a program's significance. Traditionally, standardized tests were thought to comprise the ultimate persuasive measures. When administered under controlled situations, standardized tests were said to substantiate claims of program impact. Such is not the case today. Educators are rebelling against tests which "tyrannize teachers and demoralize students" (Hein, 1975). Fortunately, alternatives to standardized tests are emerging. These alternatives are based on four assumptions.

First, direct, repeated measures of observable behavior are preferable to measuring artifacts of behavior.

Second, assessment is a process, not a product. As such the breadth of the instrument is related to the type of assessment conducted. Third, a student's performance should be compared only to himself or to that of his reference group. Fourth, the rule of parsimony should prevail. Hence, when given a choice, one should always employ the minimum number of measures which result in the maximum benefit and information.

Use Direct, Repeated Measures. The advantages of direct measures of observable behavior have been repeatedly discussed. Direct measures most closely reflect the behavior of the student in actual situations. When direct measures of behavior are obtained, the need to infer underlying meanings, motives or drives is reduced. Direct measures assess student performance in the presence or absence of specific stimuli. Behavioral artifacts are obtained when the student is presented with a contrived situation, or asked to state how he would perform in a given situation. Multiple samples of behavior are preferred because they are more reliable indicants of the actual behavior.

Direct measures of behavior provide the basic data required by all eight groups of constitutents. Hence, these measures should be monitored carefully and frequently to ensure their appropriateness. Teachers should strive to maintain complete, accurate daily records of each student's performance on target skills. These records are essential for formative evaluation purposes; they can be analyzed to determine the effects of instructional interventions and can be used to improve the programs effectiveness.

The program should also include provisions to periodically check the retention of previously learned skills and to probe for generalization of skills to to other settings. Both retention and generalization checks are useful for the teacher and program developer.

In addition, a procedure should be established to ensure periodic monitoring by the supervisor or program coordinator. Three types of checks should be conducted. First, the supervisor should document the degree to which the instructional procedures are being implemented. Second, the supervisor should verify the reliability of the measurement and recording procedures used. Third, the accuracy and completeness of daily records should be checked. These periodic checks will document program consistency and will, hopefully, counteract potential assessment hazards such as observer drift and instrument decay.

Match Breadth of Content to Assessment Phase. Because assessment is a process, a series of steps or activities should be planned, rather than a search made for an all-encompassing test. White (1980), suggests that assessment includes four phases: selection, development, implementation and refinement. Ideally, different assessment instruments would be appropriate at each level of assessment. For example, at the screening level an instrument should be selected which is broad band, (i.e., samples a wide range of behaviors). This instrument would be used for screening a large number of students or for identifying skill deficits across a wide range of behaviors. Common examples of broad-band instruments are intelligence and achievement tests. These instruments are a necessary first step in the assessment process because they identify students in need of further testing.

At the other end of the continuum -- refinement -- the testing instrument employed should, of necessity, be narrow in scope. A narrow-band test assesses a restricted set of behaviors within a skill category. Thus, a fine-grained analysis of auditory skills exemplifies a narrow-band test. Narrow-band tests are of greatest benefit to teachers, students and program developers. Their purpose is to plan and implement programs to teach specific instructional skills.

Concluding Remarks

Conceptually, the idea of matching instruments to assessment phases is obvious. The process of matching instruments to purposes should be simple. In practice, however, the reverse often occurs. According to Thurlow and Ysseldyke (1979), few guidelines are available to indicate what constitutes appropriate educational assessment. In a recent study, they surveyed the currently used assessment and decision-making practices reported by 36 Child Service Demonstration Centers. Nearly all centers were found to use all kinds of assessment instruments for all purposes. The authors were unable to detect a match between assessment phases and assessment devices. It is apparent from this study that educators need to be more careful in matching their tests to their testing purposes. It is also apparent that guidelines for assessment instrument selection are sadly needed.

Guidelines for choosing assessment instruments could be based on two lines of reasoning. In one line of reasoning, tests would be matched individually to assessment levels. Thus, for example, the Keymath may be recommended for developing instructional programs, while the Stanford Achievement Test would be used for screening. A project director using this line of reasoning would survey the available instruments and match each of them to an assessment level. The end result of this exercise would be an extensive list of tests cross-referenced with assessment levels.

A second line of reasoning would be to use assessment instruments in a number of different ways. Proponents of this theory would attempt to choose the fewest number of instruments which serve the greatest number of purposes. For example, Algozzine, (1980), advocates using PIAT results for both screening and program development. In another example, Hansen and Eaton (1978) described a method for using direct measures of reading in context at all measurement levels.

Choose appropriate comparison. Data are meaningless without anchors. Unless we have something to compare assessment results with, the data themselves are worthless. Comparisons are made possible by choosing appropriate reference points. Common points of reference used in education are 1) historical precedents, 2) people, 3) skill criterion, and 4) the individual.

Historical precedents provide an often overlooked but potentially useful source of comparison. Indeed, in some situations they are the only viable comparison. Essentially, if one appeals to an historical precedent, one cites evidence to indicate how a particular group of individuals has performed in a given situation. For example, historically, severely mentally retarded persons were automatically institutionalized. In institutions they were denied the opportunity to develop independent living skills. Due to this historical precedent, any program which improves the adaptive living skills of the severely retarded is deemed effective because it fills a void.

Historical precedents are also used to identify populations who are "at risk" for exhibiting certain problematic behaviors. Thus, if we can document that an individual with certain characteristics has a high probability of incurring learning problems and that those learning problems can be ameliorated through a specified treatment program, then claims of program effectiveness are strengthened. Many early intervention and vocational training programs depend on arguments based on historical precedents.

Norm-referenced tests are examples of reference points utilizing other people. These tests are probably the most widely used and abused measures. These tests are often inappropriate when applied to handicapped populations because they were normed on students in regular classrooms. Therefore, the comparison population is not representative of the population being tested. Additional difficulties with most norm-referenced tests are that they are often too broad band or that they test skills too

advanced for a given population. These characteristics cause norm-referenced tests to be insensitive to changes in handicapped students' performance. Suffice it to say that the concept of norm-referenced testing is valuable; it is merely the presently available tests which are unsuitable.

Criterion-referenced tests use skills as the point of reference; they compare a student's performance to some specified instructional level of skill mastery. Performance scores are expressed in terms of how closely the individual student's performance approximates the target behavior. These tests are often preferred by educators, because they concentrate on assessing an individual's skills relative to some measure of competence rather than comparing his performance to that of another student without regard to competence.

A fourth frame of reference is provided by the individual himself. This type of comparison is advocated by applied behavior analysts. In this paradigm, the individual's baseline or initial level of performance is compared to his performance during and after instruction. A functional relationship is established through the use of experimental designs (refer to Hersen & Barlow, 1976 for additional information). Individually referenced measures are recommended because they avoid the problems inherent in norm-referenced tests. Their relationship to criterion-referenced measures is presently unclear, although recently developed designs such as the changing criterion design are attempts toward resolving this concern.

Three additional types of comparisons have been suggested: normative peer data (Walker & Hops, 1976), progress monitoring (Deno & Mirkin, 1977) and systematic replications (Baer, Wolf & Risley, 1968). These methods are combinations of the other types of reference points. In the normative peer data method, an individual's performance is compared to the performance of one or more peers who are in a less restrictive environment. If the individual performs similarly to his peers, then his

behavior is comparable and the skill is sufficiently mastered. For program evaluation purposes, direct measures of a student and of his peers' performances are obtained twice; usually before and after instruction. The discrepancy between the scores at each interval are compared to determine if the student's performance has improved relative to that of his peer. If the discrepancy between the two students is reduced, then learning has occurred. The major problem with this method of program evaluation is the tendency for regression toward the mean. It appears, therefore, that this method may be most appropriate for making decisions regarding least restrictive placements.

Progress monitoring combines aspects of normative and criterion-referenced measurement. In this system, a student's performance is monitored over time on different tasks. The system is normative in that the student's rate of progress is compared to the rate of mastery exhibited by normal peers on the same set of objectives. It is criterion referenced because a set of objectives is clearly defined and specific criteria for mastery are set. The purported advantages of this technique are that the student is evaluated using direct measures of performance and he is compared to his peers, thereby facilitating decisions pertaining to his readiness for a less restrictive environment.

Systematic replications involve building a case for program effectiveness based on accumulating a number of individual replications of an instructional intervention. Thus, if it can be shown that a certain functional relationship is established across a number of different situations or across a number of students, then support is generated to substantiate program effectiveness. Reinforcement theory is one example of the power of systematic replication.

Observe the rule of parsimony. The rule of parsimony states that wherever possible, the fewest measures which provide the most data requested by the greatest numbers

of audiences should be used. Unfortunately, more people can correctly state the rule than can consistently follow it. For example, Child Service Demonstration Centers reportedly use between 3 and 39 (mean = 11.5) assessment devices (Thurlow & Ysseldyke, 1979). The wide range in the number of assessment devices is probably representative of the diversity in the number of assessment instruments used by other educational programs. Whether this diversity is necessary or only reflects the confusion evident in our present assessment practices remains to be seen. Suffice it to say, many of these assessment instruments are most likely redundant or inappropriate.

One method for reducing the number of assessment instruments used is to specify carefully the information needed to make a decision based on the intended audience and to obtain only those data. Another method would be to use the same data for different purposes. For example, Rubenstein and Nassif-Royer (1977) report a system in which state departments of education are using criterion-referenced tests for statewide assessments. Direct measures have also been advocated for use as an alternative to achievement tests for summative measures (Eaton & Lovitt, 1972).

Parsimony can be applied to methods for documenting how well students learn a skill if the measures selected are sufficiently flexible. This rule can also apply to the types of information necessary to make comparisons. Thus, the basic concerns of all eight groups of constituents can be met by developing a well-rounded child assessment system. This system should include some supporting information to provide the minimum data base necessary to respond to the concerns of all eight audiences.

Three types of supporting data should be maintained: 1) demographics, 2) cost data, and 3) measures of impact. Demographic data are used to describe the characteristics of the intended population. Cost data are used to

document staff requirements (e.g., training, instructing, analyzing), resources needed (e.g., special materials, physical space for storage and instruction) and installation vs. maintenance costs. Finally, impact data may include ratings of consumer satisfaction, indicants of interest generated through demonstration and dissemination activities and validation obtained through the results of third party evaluations, evidence of systematic replications, or testimonials of support by local, state and regional administrators. These supporting data are important for other educators, legislators and philosophers. They provide additional evidence which can be used to persuade these constituents that a program is effective and that it deserves continued support.

CONCLUSION

Persuasive educational programs provide data to their constituents which meet three basic criteria. First, they provide evidence of program effectiveness that is valid and reliable. This evidence ensures a program's credibility. Second, they provide evidence that the effect of the program is educationally significant. Significance is a function of the size of the effect, its importance and its relative cost. Third, they document the "transportability" of the program. Evidence is obtained to show that the intervention and its effects can be reproduced at other sites.

Programs which adhere to these criteria are believable. They are the types of programs sought by the Joint Dissemination Review Panel (Tallmadge, 1977). They are indeed exemplary model programs.

REFERENCE LIST

Adelman, H. S. & Taylor, L. Initial Psychoeducational assessment and related consultation. Learning Disabilities Quarterly, 1979, 2, 52-64.

Algozzine, R., & McGraw, K. Diagnostic Testing in Mathematics: An Extension of the PIAT. Teaching Exceptional Children, 1980, 12, 71-77.

Baer, D., Wolf, M. M., & Risley, T. R. Some current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis, 1968, 1, 91-97.

Deno, S. L., & Mirkin, P. Data-based program modification. Reston, VA: Council for Exceptional Children, 1977.

Hansen, C. L. & Eaton, M. D. Reading. In Haring, N. G., Lovitt, T. C., Eaton, M. D., and Hansen, C. L. (Eds.) The fourth R: Research in the classroom. Columbus: Charles E. Merrill, 1978.

Eaton, M., & Lovitt, T. C. Achievement tests vs. direct and daily measurement. In G. Semb (Ed.), Behavior analysis and education--1972. Lawrence, KS: University of Kansas Press, 1972.

Hein, G. E., Standardized testing: Reform is not enough! In Perrone, V., Cohen, M. D., & Martin L. P. (Eds). Testing and Evaluation: New Views. Washington DC: Association for Childhood Education International, 1975.

Hersen, M., & Barlow, D. H. Single case experimental designs: Strategies for studying behavior change. New York: Pergamon Press, 1976.

Rubenstein, S. A., & Nassif-Royer, P. The outcomes of statewide assessment: Implications for curriculum evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1977. (ERIC Document Reproduction Service No. ED 141-394).

Tallmadge, G. K. The joint dissemination review panel ideabook. Mountain View, CA: RMC Research Corp., 1977. (Available from U.S. Government Printing Office, Washington, DC.)

Thurlow, M. L. & Ysseldyke, J. E. Current assessment and decision making practices in model LD programs. Learning Disabilities Quarterly, 1979, 2 (4), 15-24.

Walker, H. & Hops, H. Use of normative peer data as a standard for evaluating classroom treatment effects. Journal of Applied Behavior Analysis, 1976, 9, 159-168.

White, O. Basic Considerations in Child Assessment: Not Quite Everything You Wanted to Know ... and More. In C. Hansen (Ed.), N. Haring, Series Editor. Child assessment: The process & the product. Seattle: Program Development Assistance System, University of Washington, 1980.

Wolf, M. M. Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. Journal of Applied Behavior Analysis, 1978, 11, 203-214.

Participants

Bob Algozzine
G315 Norman Hall
University of Florida
Gainsville, FL 32611

Bill Banaghan
Project HELP
2430 Stanwell Drive, Suite 160
Concord, CA 94520

John Bjorklund
Project Explore
St. Paul Public Schools
Mechanic Arts High School Building
97 E Central
St. Paul, MN 55101

Jimmy Clark
Social Behavioral Survival Program
Center on Human Development
College of Education, Room 206
University of Oregon
Eugene, OR 97403

Melvin Cohen
Augmentative Communication Model Program
Loma Linda University Medical Center
Department of Speech and Language Development
University Arts Building, Suite 104
Loma Linda, CA 92350

George Culp
Project PRISM
Portland High School
95 High Street
Portland, CT 06480

John Davidson
The Adaptive Learning Environments Model: A Mainstreaming Program of Mildly Handicapped Children
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, PA 15261

Dave Dole
Project PRISM
Portland High School
95 High Street
Portland, CT 06480

Leslie Ellis
Project REACH
San Luis Valley Board of Cooperative Services
22nd and San Juan
Alamosa, CO 81101

Vicki Engler
Project REACH
San Luis Valley Board of Cooperative Services
22nd and San Juan
Alamosa, CO 81101

Virginia Evey
Project CAST
Charles County Board of Education
Pomombey Annex
La Plata, MD 20646

Bataan Faigao
Foxfire Model Project
Magnolia Star Route
Nederland, CO 80466

Cheryl Hansen
Program Development Assistance System
University District Bldg. JD-11
1107 NE 45th - Suite 330
Seattle, WA 98105

Richard Kissane
Life Adjustment and Employment Preparation for Special Students
BOCES of Nassau County
The Salisbury Center
Valentine Road and Plains Road
Westbury, NY 11590

Peppy G. Linden
A Model Program of Early Education for the Cerebral Palsied Child in a Rural Setting
University of Virginia
Department of Pediatrics
Jefferson Park Avenue
Charlottesville, VA 22903

Corrine McGuigan
University of Idaho
College of Education
Moscow, ID 83843

Martin Miller
Project TIDE
1200 Waters Place
Room B-1077
Bronx, NY 10461

Janet Morrison
Project Mainstream
Brockton Public Schools
43 Crescent Street
Brockton, MA 02401

Marge Patten
Project KEYE
Klein Independent School District
Resource Service Department
7200 Spring-Cyprus Road
Spring, TX 77379

Irene Potosky
Project CAST
Charles County Board of Education
Pomombey Annex
La Plate, MD 20646

Bruce Richards
Community Teaching Homes
School for Contemporary Education, Inc.
623 South Pickett Street
Alexandria, VA 22304

Irwin Rosenthal
Learning Opportunities Center for Special Community College Students
Department of Student Services
Kingsborough Community College
2001 Oriental Boulevard
Brooklyn, NY 11235

Charity Rowland
Engineering Process-Oriented Educational Programming for SPH Adolescents
Bureau of Child Research
University of Kansas
Parsons Research Center, Box 738
Parsons, KS 67357

Dana Simons
Project SEED
Dallas Independent School District
Special Education
3700 Ross Avenue
Dallas, TX 75204

Owen White
Program Development Assistance System
University District Bldg. JD-11
1107 NE 45th - Suite 330
Seattle, WA 98105