DOCUMENT RESUME

ED 118 649 95 TM 005 145

AUTHOR Sanders, James R., Ed.; Sachse, Thomas P., Ed.
TITLE Problems and Potentials of Applied Performance Testing. Proceedings of the National Conference on the Future of Applied Performance Testing.
INSTITUTION Northwest Regional Educational Lab., Portland, Oreg.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE Dec 75
NOTE 143p.

EDRS PRICE MF-$0.83 HC-$7.35 Plus Postage
DESCRIPTORS Clearinghouses; *Conference Reports; Elementary Secondary Education; Evaluation; Guidelines; Instructional Materials; Military Training; *Performance Tests; Research Needs; *State of the Art Reviews; Student Evaluation; Teacher Education; Testing Problems; Training
IDENTIFIERS Clearinghouse on Applied Performance Testing; Elementary Secondary Education Act Title V; ESEA Title V

ABSTRACT

The purpose of this conference was to share the information gathered by the Clearinghouse for Applied Performance Testing (CAPT) such as information on performance testing that could be used in public schools, and secondly, to discuss problems that must be solved, issues that should be addressed, and additional research and development needed in the area of Applied Performance Testing (APT). The presentations by the Clearinghouse dealt with the state of the art of APT, an overview of Clearinghouse activities, instructional materials developed on APT, and guidelines for the evaluation of APT materials and procedures. The invited address by Saul Livisky was presented next. This was followed by small group discussion reports on problems, issues, and needed research development in APT. In the next section individual papers are presented discussing these problems, issues, etc. Appendices contain participants in this 1975 conference, handouts accompanying the invited address, and guidelines for the evaluation of APT materials and procedures. (RC)


DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

The work presented or reported herein was performed pursuant to a Grant from the U. S. Office of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the U. S. Office of Education, and no official endorsement by the U. S. Office of Education should be inferred.

PROBLEMS AND POTENTIALS OF APPLIED PERFORMANCE TESTING

PROCEEDINGS OF THE NATIONAL CONFERENCE ON THE FUTURE OF APPLIED PERFORMANCE TESTING

Edited by
James R. Sanders[1] and Thomas P. Sachse
Clearinghouse for Applied Performance Testing,
A Project Supported by Funds from the
U.S. Office of Education, ESEA Title V 505

Northwest Regional Educational Laboratory
710 S.W. Second Avenue
Portland, Oregon 97204

December 1975

[1]Dr. Sanders is presently Associate Professor at The Evaluation Center, Western Michigan University.

The National Conference on the Future of Applied Performance Testing, cosponsored by the Clearinghouse for Applied Performance Testing at the Northwest Regional Educational Laboratory and the National Council on Measurement in Education, was held at the Washington, D.C., Hilton on March 30, 1975.

ACKNOWLEDGEMENTS

The editors wish to thank the many people who participated in the National Conference on the Future of Applied Performance Testing and who made this publication possible. We are grateful for the continued support of ESEA Title V, Section 505, and its staff at the U.S. Office of Education, particularly Dr. Thomas Johns. This support has been enhanced by the productive participation in the Clearinghouse from representatives of four participating states: Dr. Gordon B. Ensign, Jr., Dr. James C. Impara, Mrs. Pauline Leet, and Dr. Janet I. Sumida. Without the support of these people, the Clearinghouse would not have been possible.

Members of the staff of the Assessment Program at the Northwest Regional Educational Laboratory who provided the backing needed to make the Conference a success include: Robert Keysor, Dean Nafziger, James Olsen and Brent Thompson. We wish to thank our colleagues for their valuable assistance.

We are deeply indebted to Peggy Hootstein and Julie Stange who have stood by us during some trying times while the Conference was planned and conducted and who carefully prepared materials used during the Conference. Further, Vicki Spandel has provided excellent editorial support during the preparation of this report.

Finally, we express our sincere appreciation to the National Council on Measurement in Education and to the many professionals who participated in the Conference and for the papers produced by the Conference presenters. The editors, however, accept all responsibility for any shortcomings of the Proceedings.

TABLE OF CONTENTS

INTRODUCTION

James R. Sanders
Clearinghouse for Applied Performance Testing

The Clearinghouse for Applied Performance Testing (CAPT) and the National Council of Measurement in Education (NCME) would like to welcome you to the National Conference on the Future of Applied Performance Testing. Basically, we see this as a working conference rather than a didactic session. The purpose of the conference is two-fold. First, we wish to share with you information that the Clearinghouse has gathered over the last nine months, such as information on performance testing that could be used in public schools. Second, we wish to provide an opportunity for discussion of problems that must be solved, issues that should be addressed, and additional research and development needed in the area of applied performance testing. We look to you, the audience, for direction to guide that discussion.

Everyone at the conference is involved in applied performance testing at some level. In essence, we have most of the experienced and knowledgeable persons in the field gathered in this room. This has significant implications in terms of what can be achieved. This conference does not represent an isolated effort. We hope to lay out specific goals for ourselves--and for others in this field--and then reconvene next year to see how well we have achieved those goals and to plan the next steps.

This conference is to be structured in the following way: first, we would like to describe the operation of the Clearinghouse and to share some of the information we have collected. The best way to do that is to ask members of the Clearinghouse Policy Board to talk with

you briefly about some activities for which they have taken responsibility. Second, we have invited Mr. Saul Lavisky from HumRRO to address the group this afternoon. HumRRO is one of the oldtimers in terms of applied performance testing and training. Because of his long-term experience, Mr. Lavisky can present a perspective that many of us have not had the opportunity to develop, and we appreciate his willingness to share that with us.

Following Mr. Lavisky's remarks, we will break into smaller groups to address discussion questions the conference staff have laid out. Groups will be formed on the basis of professional role, with administrators in one group, curriculum specialists in another, and measurement and evaluation specialists in a third.

The small group sessions will be task oriented. We have asked each group to address three questions from their particular perspectives as representatives of a specific profession. The first is, What problems are involved in the development or use of applied performance testing in public schools? Second, What issues arise when applied performance tests are considered for use in public schools? Finally, What research and development efforts are needed in the area of applied performance testing? We have asked one participant from each group to serve as a recorder and provide the larger group a summary of the small group's discussion of each question. We will reconvene later this afternoon to hear those reports.

For the evening session we have asked four discussants, all people who are extensively involved in applied performance testing, to share with us their thoughts about what direction applied performance testing is now taking. The four discussants are Joseph Boyd from the Educational Testing Service; Hulda Grobman from the University of Illinois, College

of Medicine; William Osborn, another HumRRO representative; and Ruth Nickse from Syracuse University Research Corporation. Each of these four people has agreed to use his or her expertise in helping us effectively address the topic to be covered during the course of this conference.

Members of the Clearinghouse Policy Board have agreed to summarize for you the work that the Clearinghouse on Applied Performance Testing has done this past year. Let me provide a brief background for their remarks. The Clearinghouse was established in July, 1974, through a grant from Title V, Section 505 of ESEA to four participating states: Hawaii, Oregon, Pennsylvania and Washington. Subsequently, two other projects were added to the Clearinghouse effort: one by the U. S. Office of Education, Office of Planning, Budgeting and Evaluation, to collect and evaluate measures of functional adult literacy; and one, initiated through the Department of Defense, to search for occupational certification measures.

The Clearinghouse Board has asked the Northwest Regional Educational Laboratory to oversee day-to-day Clearinghouse operations and we have a Clearinghouse staff at the Laboratory; Mr. Thomas Sachse is here representing that staff.

The Policy Board members are: Dr. Janet I. Sumida, Director of Statewide Assessment at the Hawaii Department of Education and Project Administrator for the Clearinghouse project; Dr. James Impara, Director of Statewide Assessment at the Oregon Department of Education; Mrs. Pauline Leet, Director of the Bureau of Curriculum Services at the Pennsylvania Department of Education; and Gordon B. Ensign, Jr., Supervisor of Program Evaluation with the Washington Superintendent of Public Instruction's office. Since each Policy Board member has undertaken

specific tasks on applied performance testing, I will let them now describe for you what they have been doing.

II.

REPORT BY THE

CLEARINGHOUSE FOR APPLIED PERFORMANCE TESTING

APPLIED PERFORMANCE TESTING--THE STATE OF THE ART

James C. Impara
Oregon State Department of Education

Applied performance testing is not a new concept. In fact, applied performance is probably one of the oldest forms of testing known. However, it is difficult to find measures which have been "standardized" so that the administration, scoring and interpretation are reliable. Exceptions to this occur in a number of military settings and in some vocational settings, but it is not the case that measures are available for the more "mundane" activities (performances) each of us encounters on a regular basis.

In an attempt to learn what currently exists in the field of applied performance testing, a literature search for tests or informal papers was conducted. In addition to the literature search a survey was conducted to obtain materials relevant to applied performance testing.

The literature search focused mainly on publications and projects developed during the last five years. It was learned, however, that military sources of information required more extensive research since performance testing had been employed by the military since World War II. Searches of computer information bases were conducted to reveal additional sources of information.

Results of the searches varied widely (even within the same system) depending on the search strategy and descriptors used. This variability stemmed from the fact that descriptors used within the systems did not correspond to current notions of performance. This proved a complex problem. Not only did descriptors fail to match our descriptions (making

access difficult) but descriptors assigned when documents were entered into the system were couched in a dated vocabulary.

To aid in the search for materials, a subcontract was given to Adrian Van Mondfrans of Bringham Young University. The BYU staff completed a literature survey as well as a field survey of applied performance assessment activity. The literature survey netted 350 annotated references. Many of these references duplicated the present Clearinghouse materials; however, there were enough new references to convince the Clearinghouse staff that the external search activity had been beneficial to the project.

The field survey employed a questionnaire sent to 600 individuals throughout the country. In this survey special emphasis was placed on determining the need for an availability of instructional materials and measures for applied performance assessment. Unfortunately, the return rate of this questionnaire was quite low--perhaps because it had been sent just prior to Christmas 1974. However, a follow-up study retrieved some additional data.

By winter of 1974, many projects had been identified in the field and approximately 30 major centers of activity for field activity were noted. In an attempt to collect current information about new developments in the field, the policy board and staff visited these projects. Although some projects were not as deeply involved in testing as originally believed, these site-visits proved beneficial since most projects had materials and references that were of great utility to the Clearinghouse and Clearinghouse users.

The data collection activities described above portray somewhat the state of the art in applied performance testing. As might be expected, applied performance testing is well developed in subject matter areas in

which the product of the education requires the ability to perform. Occupational fields, such as carpentry, mechanics, clerical skills and masonry rely on both performance tests and complementary paper and pencil tests to certify occupational competency. Professional occupations--especially the medical arts and teaching--have been very active in using performance training and testing. The military and private industry have also used performance testing extensively.

Simulation is a well-developed facet of applied performance testing. Business and the medical arts are proficient users of simulation and gaming techniques. Simulation has some distinct advantages over performance testing, including reduced cost, increased sampling of behavior, and the possibility for variation while maintaining standardization.

Although traditional public school content areas often lack applied performance testing devices, increased interest and development in basic skills assessment will soon change this. A growing desire for assessment of school subject matter in terms of life skills will require additional measures of an applied performance nature. As a result of Clearinghouse activities, some technical instruments for measurement of public school content areas are becoming available.

The unwillingness of some developers to share their ideas and products has been a major problem for the Clearinghouse. This unwillingness seems to stem from two different points of view. The first is that the producer of the measures does not feel that the measures are ready to be released for wide-scale use before further development. This problem might be classified as the avoidance of potential embarrassment because of known or expected flaws in the development of the measure. Whenever the Clearinghouse is aware of such a circumstance, it has offered to guarantee the author's anonymity as well as provide feedback on the

materials to the author if desired; the Clearinghouse is very concerned that the developers of new, incomplete materials be protected from potential embarrassment because of uncorrected flaws or errors. The second point of view stems from the unwillingness of certain groups to participate because they have expended large sums of money or large amounts of time (or both) and do not wish their materials to be distributed to those who have not participated in the development.

It is hoped that as the Clearinghouse grows, both with respect to the collection of material and with respect to establishing trust and credibility that the reticence shown by some who are unwilling to share will be relieved and the state of the art can grow even further.

AN OVERVIEW OF CAPT ACTIVITIES

Thomas P. Sachse
Clearinghouse for Applied Performance Testing

The Clearinghouse for Applied Performance Testing (CAPT) was cooperatively conceived, proposed and undertaken by representatives of NWREL and the four member states. This group, collectively designated the CAPT Policy Board, directs Clearinghouse operations. It should be noted that the Policy Board has done an excellent job of delineating the tasks necessary to operationalize a clearinghouse of this type. Let me now describe for you the activities that have shaped the present status of CAPT.

Throughout the short life of the project, the CAPT proposal has provided clear directions for developing a functional clearinghouse. In writing the proposal, the Policy Board took care to provide means for accomplishing six main objectives:

1. Collection of applied performance testing materials.
2. Formation of a consumer audience interested or involved in applied performance testing.
3. Dissemination of materials of vital concern to potential users.
4. Development of instructional materials on applied performance testing.
5. Development of criteria for evaluating applied performance instruments.
6. Evaluation of the CAPT project.

Members of the Policy Board are here today to discuss their roles in the completion of these objectives. My overview of the CAPT project is intended to complement their remarks.

Dr. Impara has already described the results of important collection activities undertaken this past year. I will now delineate the tasks that led to those results. Let me first mention that many CAPT activities were conducted for multiple purposes. For example, one of the first activities was to publicly announce the establishment of CAPT and to solicit information about persons or projects in the field. This activity provided CAPT its first materials and began the formation of our consumer audience. These releases were sent to a variety of educational journals and other informational publications.

Numerous letters soliciting applied performance testing materials were sent to workers in the field. Their responses provided additional materials to CAPT and further expanded the consumer audience. Many researchers in the field were identified through another collection effort--namely, literature searches. During the past year, CAPT has conducted a literature search at the NWREL Information Center, four different computer information-based searches and, as Dr. Impara noted, has contracted with BYU to conduct an independent literature search and field survey.

Surprisingly, no search entirely duplicated previous efforts. As a result of these searches, additional applied performance testing materials were contributed to CAPT; and additional workers in the field were identified.

Collecting materials and forming a CAPT consumer audience are ongoing activities. CAPT receives, daily, requests from persons wishing to be put on the mailing list, and regular contributions of testing materials.

By November and December of 1974, major centers of activity had been identified, and plans were made to visit these projects to collect information and materials. CAPT has received continued interest and

support from those project personnel; more than half of our conference discussants and reporters are from agencies visited by CAPT.

This National Conference is primarily viewed as a planning and dissemination activity. It is expected--and hoped--that our discussions will net new sources of information for CAPT.

The formation of a consumer audience has been closely tied with collection efforts. As important projects or products in the field are identified, CAPT contacts individuals responsible for development, as well as other interested persons, so that through information sharing all can benefit from others' endeavors. Persons who contributed materials to CAPT are offered, in exchange, an equivalent number of duplicated materials at no cost. Our concern in making the Clearinghouse effective in dissemination as well as collection has resulted in an active consumer audience.

A number of important tasks must be completed prior to dissemination. When materials are first received, they are screened to determine their appropriateness for inclusion in our collection. Because applied performance testing is a large umbrella under which many testing devices and materials fall, this task may seem unnecessary. On the contrary, however, a surprising number of contributions bear little relation to applied performance testing. They are occasionally included, however, because even materials that seem only tangential to the field are often requested by our consumer audience.

CAPT was formed to collect testing materials for use at the public school level; however, most contributions fall outside this domain. The relative newness of the field and the use of performance settings (occupational and adult education) are two factors that encourage development of applied performance testing in non-traditional public school subject

matter areas. Recent demand for applied performance tests in public schools will shift the developmental emphasis, and CAPT will be ready to provide assistance where necessary.

Once materials have been screened, they are referenced and catalogued for user access. Subject matter, target population, availability, grade level and testing mode are but a few of the variables by which the materials are classified.

Availability is an area of particular concern to CAPT. For many reasons, some of the finest materials collected are unavailable to CAPT in quantities adequate for dissemination. If such materials are available from a specified source, CAPT tries to provide users ordering information.

Unfortunately, public school educators are not funded to obtain commercially available materials. Many excellent products--capable of providing much valuable information--are still in developmental stages. The interest in applied performance testing has now developed far beyond the field's technical or financial capability to respond.

CAPT is presently annotating screened and catalogued materials to provide users the kind of summary information they need in requesting CAPT materials. Having to select materials strictly on the basis of title, author, institution, date and number of pages is simply unsatisfactory.

Dissemination of collected materials and information on applied performance testing has been and remains the ultimate goal of CAPT. CAPT was established to provide materials to those with an expressed need; filling expressed and perceived needs is a constantly expanding activity. During the first months of CAPT, the emphasis was on collection, now we must "deliver the goods" to interested parties. Although

collection and dissemination are complementary, ongoing activities, shifts in emphasis do occur over time.

At the inception of the CAPT project, an informational brochure was distributed to (1) announce the establishment of CAPT and (2) solicit applied performance materials. That brochure was also designed to introduce people to a concept in measurement that they may have overlooked in relation to public school testing.

The CAPT Newsletter, published bimonthly, has been useful in keeping readers informed of new developments in the field, CAPT activities and projects and publications relating to applied performance testing. The January CAPT Newsletter included a list of References Related to Applied Performance Testing. This document provided readers an initial look at the then current status of publications in the field. Ordering information was included. The annotated bibliography of CAPT resources--an updated version of that list of references--will be released in May, as will The Synthesis Survey of Applied Performance Materials, a state-of-the-art document of all references encountered by CAPT.

Dissemination of CAPT materials made available through the "References Related" document has been constant since the January Newsletter was issued. In addition, CAPT has responded to inquiries for help in applied performance testing and to requests for assistance in statewide assessment. Individual requests for CAPT assistance are handled by the CAPT staff or by member state representatives to CAPT. These requests vary greatly and many specialized needs cannot be met. In the event that a testing device is not currently available for the subject matter or audience, materials or devices that can be adapted are recommended. In the event that CAPT cannot meet an individual's need, the request is filed and reactivated when new relevant materials are collected.

CAPT has also provided support to various statewide assessment activities. The State of Hawaii is currently pilot-testing applied performance exercises for use in statewide assessment. Oregon and Pennsylvania have indicated a need for applied performance measures of citizenship, and CAPT is assisting in the development of testing materials for this important subject matter. CAPT has provided applied performance materials to non-member states as well.

Through a variety of approaches, CAPT has attempted to acquire a national scope. For example, publicity releases were issued to national publications and regional publications outside of member states. CAPT has received contributions and requests to be put on the mailing list from persons in almost every state. The CAPT audience is, however, concentrated in member states, because of extensive publicity provided through member state publications, the influence of member state representatives, and dissemination policies for member states--CAPT materials are free to agencies within member states.

The reasons for seeking a national scope are (1) to increase interest in and contributions to CAPT holdings, (2) to decrease duplication of development efforts, and (3) to determine appropriate future activities for CAPT and those interested in promoting the field of applied performance testing.

This National Conference represents a culmination of efforts to achieve national scope. Because the National Council of Measurement in Education (NCME) is co-sponsoring the Conference, announcements of the Conference were sent to the entire NCME mailing list.

CAPT is responsive to problems in the field of educational measurement and has attempted, to the extent that available resources permit, to relate applied performance testing to larger educational goals. This

past spring, CAPT invited three well-known measurement specialists to discuss different approaches to measuring student competencies. Dr. Robert Ebel, of Michigan State University, discussed traditional norm-referenced testing; Dr. W. James Popham, of the University of California at Los Angeles, dealt with domain and criterion-referenced measurement; and Dr. William McClelland, of the Human Resources Research Organizat ..., reported on applied performance testing.

No consensus concerning the most effective approach to competency-based testing was reached. The group noted a need for a more adequate definition of the term "competency-based measurement," and proposed specific approaches to competency-based assessment in Oregon. The comments were made in the context of legislative mandates for competency-based measurement in Oregon.

CAPT has been represented in the NCME Task Force on Competency Measurement. This Task Force has been asked to identify major issues concerning competency-based measurement, and to suggest strategies and directions for future research.

The Clearinghouse Policy Board is concerned with advancing the field of applied performance testing, and to this end has begun developing various facets of the field. Each state representative has taken responsibility for one or more CAPT activities in addition to Policy Board direction, site visitations, and state responsibilities.

DEVELOPING INSERVICE EDUCATIONAL MATERIALS FOR APPLIED PERFORMANCE TESTING

William Gauthier, Jr., Bucknell University,
Pauline Leet, Pennsylvania Department of Education and
Hugh F. McKeegan, Bucknell University

Introduction

One objective of the Clearinghouse for Applied Performance Testing is the production and evaluation of instructional materials on "the definition, attributes, development and use of applied performance procedures and materials for student assessment." Discussion of this work will complement the literature search and field survey conducted by Richard Kay and others (1975), and add to the resources already available from NWREL, ERIC, HumRRO, and other relevant information sources. The primary emphasis of this work is on the "production of new instructional materials relevant to consumer needs."

Rationale

Many important outcomes of elementary and secondary education are defined through an "if x then y" kind of relationship. In other words, if a student masters skill "x," he has mastered, or probably will master, skill "y." Certain "x" skills--such as reading, writing, computing and speaking--can be assessed directly; other kinds of cognitive competencies, such as the degree of mastery of a secondary level course in history or literature, are usually evaluated by sampling the behaviors which the course is purported to develop. Whether one uses a criterion or norm-referenced approach, "x" type learnings can be assessed rather reliably for the purposes of the school using a variety of direct and indirect measurement techniques. Traditionally the "y" kinds of outcomes (e.g.,

good citizenship, work habits, social responsibility) are believed to arise from command of subject matter skills and immersion in the micro-society represented in the school.

Applied performance testing requires educators to re-examine the extent to which their assessment of type "x" outcomes is reliable and valid when it occurs in a context, either real or simulated, representative of the macro-society. APT also demands a searching analysis of "y" type outcomes to determine (a) the extent to which they can be operationalized and assessed directly, (b) the degree to which it can be logically inferred that mastery of "x" type outcomes will provide a basis for appropriate behavior in "y" type situations or conditions. Despite our best efforts in these analyses, there will always be relative uncertainty about the behavior individual graduates will exhibit in complex real-life situations. Personality factors, attitudes; the nature of the problem, the situational context in which the problem is presented and the degree of originality in generating problem-solving strategies are but a few of the factors that affect real-life performance. Further, to paraphrase Margaret Meade, "We must often teach for what we don't know yet," for a future that is undefined, schools must stress analytical and problem-solving skills and procedures that will have general applicability. Nevertheless, education depends extensively on reducing uncertainty in behavior, and APT can contribute significantly to this effort by expanding and improving the assessment devices and procedures used in schools.

As this discussion indicates, the development of teacher inservice materials for APT involves curricular as well as measurement and evaluation considerations. The survey conducted by Kay et al., together with a variety of observations made by preservice and inservice teachers, suggests the extent of need in the measurement and evaluation area: it

can only be described as enormous. Only a small proportion of teachers appear to have adequate command of measurement theory, classical test concepts, or newer criterion-based approaches. The inservice materials to be developed for APT will assume a basic knowledge of elementary concepts in traditional tests and measurement, but most schools contemplating the use of APT will need to structure their inservice activities so that teacher competence in these prerequisite areas is assured. To this end, the materials will include references to selected materials and programs already available in the general area of testing and evaluation with particular emphasis on criterion-referenced measurement. Either voluntary or mandated use of APT will require individual teachers' participation in analysis and re-analysis of curriculum priorities, as well as the appropriate use of a variety of applied performance tests, indices, and observations. The inservice materials will be designed to contribute to effective participation in curriculum decisions impinging on APT and to the effective and appropriate use of APT procedures.

Improving the competence of individual teachers--either preservice or inservice--while certainly necessary, will not ensure that applied performance techniques are appropriately used in schools. In summarizing the research on educational innovations, Spady concludes that "the failure of many if not most innovations lies in the failure of schools to implement them adequately." Other observers, particularly those involved in current developments in Oregon, emphasize the enormity of the institutional change involved in developing APT programs. Developing either teacher competence alone or administrator competence alone can only lead to a great deal of personal frustration and institutional fragmentation in the implementation of any sizable innovation. Thus, it would appear that there should be a sequenced development of the competencies of

decision makers at both the classroom and district levels if APT is to be more than another innovative fad. The nature of the decisions to be made and the kinds of information to be collected and processed are quite different at the classroom and institutional levels. Administrators must concern themselves with such topics as determining needs, conducting discrepancy analysis, establishing priorities, securing staff commitment, developing goals and objectives, and implementing, evaluating and refining pilot programs. And while they must have a cognitive understanding of APT concepts and procedures similar to that of the teacher, administrators must also attend to all procedures and constraints involved in bringing about viable and defensible institutional change. To meet these diverse needs, the inservice materials being developed are two distinct but coordinated units. The first would focus on the informational needs of the classroom teacher in implementing the specifics of an APT program. The second would attempt to meet the informational needs of department chairmen, coordinators, principals, and superintendents who must establish institutional parameters and priorities for APT.

Constraints

The major constraints involved in developing the inservice materials center on (a) the state of the art in applied performance testing, (b) the availability of appropriate examples of applied performance testing, and (c) the time frame in which materials must be completed. Quite sophisticated applied performance strategies have been developed--particularly by HumRRO--for use in military training contexts, and tests of an applied performance type have been developed by certain government agencies, industries and vocational education institutions. Applicability of these materials to the kinds of tasks in which public schools

desire applied performance assessment may be rather limited. While some elementary and secondary schools have developed applied performance tests or use applied performance procedures, examples of these kinds of approaches are not readily available to schools--especially teacher training institutions. The work of the Clearinghouse for Applied Performance Testing in collecting fugitive materials and in encouraging their further refinement and standardization should do much to alleviate this problem. Contractual deadlines and requirements of the funding agency are such, however, that the collection of materials and the development of inservice materials described here must be completed by the end of this fiscal year. The nature of the task has been defined, literature searches have been conducted, and preliminary outlines have been prepared. Nevertheless, the products, while they should prove useful in the preservice and inservice education of teachers and administrators, must also be considered as curriculum materials subject to formative evaluation and further revision.

Description

The inservice materials will comprise two units each, a discussion guide and references. One unit will be designed for the preservice and inservice education of teachers and will include information on the definitions of APT, appropriate and inappropriate curricular uses of APT, constraints in use, and procedures for development and evaluation of applied performance tests. The second unit will focus on administrative and institutional concerns in implementing APT programs and will include components on needs analysis, systems development, pilot testing of APT based curricular and instructional systems, and formative and summative evaluations of APT programs.

Each unit will be tested in the developmental stage with small samples of preservice and inservice teachers and administrators, and their responses used as a guide to revision and improvement.

Persons who can offer suggestions regarding the development of the inservice materials or references to extant materials relating to the project are encouraged to contact developers c/o The Department of Education, Bucknell University, Lewisburg, Pa. 17837.

GUIDELINES FOR EVALUATION OF APPLIED PERFORMANCE TEST MATERIALS AND PROCEDURES[1]

Janet I. Sumida
Hawaii State Department of Education

Need for the Guidelines

As the Clearinghouse for Applied Performance Testing (CAPT) collects, processes, and disseminates applied performance test materials, some preliminary screening of the materials is necessary to ensure quality control. Guidelines for systematic evaluation and screening of the materials must be established and publicized so that users of the test materials can be selective.

In developing new test materials, CAPT must also be guided by criteria for evaluating the adequacy of the materials. Other test developers may also find it useful to have a set of established, accepted guidelines to which they can conveniently refer.

Purpose of the Guidelines

The guidelines are proposed primarily for use in evaluating the adequacy of applied performance tests. However, they may also be of help to test developers who must also be aware of the criteria for determining the adequacy of applied performance tests. The guidelines do not provide specific procedures for developing applied performance tests; they are, according to Osborn, not "how-to-do-it"[2] guidelines for test

---

[1]The complete set of guidelines are provided in Appendix C of this document.

[2]Quoted phrase from William Osborn's paper on review of the first draft of the proposed guidelines, March 1975.

developers, but rather a set of criteria for assessing whether they "have done it."[3] They are intended essentially as a means of ensuring quality control of applied performance test materials and procedures.

Ongoing Review and Updating of the Guidelines

At the K-12 school level, developmental work in the area of applied performance testing is relatively new. Technical guidelines for development and evaluation of applied performance test materials and procedures have not been formally developed, studied, or written about as extensively as those for other areas of development and measurement. Those guidelines that have been compiled so far have been (a) "borrowed" wherever appropriate from literature pertaining to traditional testing, or (b) newly developed, based on current CAPT staff experiences in the area of applied performance testing. In view of the way in which guidelines were compiled, it was necessary that they be initially reviewed by test and measurement experts.

The following criteria were proposed to the initial reviewers in their consideration of the newly compiled guidelines:

1. Communicability of guideline statements. Is there a need for additional details and further clarity?
2. Technical soundness. Are the guidelines credible, based on experience and available information?
3. Usefulness. Is the guidelines' applicability potentially broad in scope?
4. Relevance. Do the guidelines serve to fill a critical gap?
5. Updatedness. Are the guidelines consistent with current developments in the area of applied performance testing?

[3]Quoted phrase from William Osborn's paper on review of the first draft of the proposed guidelines, March 1975.

The guidelines are subject to refinement and updating; additional review and input will be solicited as they are more widely disseminated for trial use. CAPT personnel would appreciate some discussion on the guidelines during today's small group sessions. Input from the initial group of reviewers acknowledged on your copy of the guidelines has been most helpful.

How to use the Guidelines

Although "applied performance testing" has been defined for CAPT purposes, identifying the tasks of different age groups--such as fourth graders, eighth graders, and eleventh graders--as "applied performance" can be a problem. We usually associate vocational or on-the-job competencies with adults, but it appears necessary to view school age youngsters' competencies differently. Do we view students' competencies as preprequisites to on-the-job or out-in-the-world adult survival competencies? Perhaps schools can only provide indirect, inferential evidence that pupils are likely to behave completely because they possess the essential prerequisites to out-in-the-world and on-the-job competencies.

When not equated with vocational competencies, measurement of students' competencies must be based on test items that differ largely from those used in directly measuring vocational competencies. For example, we would not subject a fourth grader to a test of special stenographic skills but more appropriately to what may constitute a set of prerequisite stenographic skills such as ability to organize, to carry on a telephone conversation, to alphabetize, or to greet visitors. Such prerequisites to out-in-the-world or on-the-job competencies tend to be general and applicable to many vocational areas.

We may want to further break down applied performance according to the way Ebel proposes to identify competencies: (1) cognitive, (2) physical, and (3) personal. According to Ebel, cognitive competency results from the assimilation of useful information to form a structure of knowledge and understanding; physical competency is a result of natural endowments developed by practice; and personal competency is a result of experience, imitation and adaptive behavior modification. Such differentiated competencies could represent the kinds of prerequisite competencies we speak of in relation to applied performance testing for K-12 students.

In developing Hawaii's first statewide basic skills assessment package in the area of reading, we have attempted to identify those reading competencies that facilitate further learning and communication for the student. We have identified performance indicators to include the following:

1. Understands meaning of words, word phrases, and word relationships.
2. Demonstrates a positive attitude toward reading; reads a variety of materials (including narrative, graphs, tables and charts) for various purposes.
3. Locates and uses reading sources effectively.
4. Follows written directions.
5. Gets the main idea and supportive details from a reading selection.
6. Reads critically.

Therefore, we have put together a test package that appears to differ from a traditional reading test. Hawaii's reading test package leans toward applied performance testing. Because the reading test package makeup is somewhat unusual, it has even been recently suggested by certain local developers of reading materials that we identify our reading

assessment package by the set of performance indicators rather than identify it strictly as a traditional reading test. We would like to consider our reading assessment package as including a large part of what is traditionally covered in a reading test as well as additional applied performance material.

Instruments for the measurement of students' prerequisite competencies will have to be evaluated for adequacy according to criteria that may be represented by some traditional testing guidelines. Instruments for the measurement of occupational competencies, on the other hand, will have to be evaluated according to less traditional guidelines.

The present proposed set of guidelines therefore consists of criteria that may be used for testing of (1) general, prerequisite competencies, and (2) occupational competencies. When students' general prerequisite competencies are to be included as applied performance, we would have to accept applied performance testing as including a wide range of situations. Guidelines would then have to be viewed by users as applicable to many different test materials and procedures. Users must be selective in the application of guidelines for evaluation of unique instruments.

It has been necessary to discuss the nature of applied performance and related test content to arrive at a common understanding about the basis for the selection of guidelines presently proposed for your review.

We have made a beginning in the search and development of guidelines. Your continued involvement and contributions are most appreciated.

III.

INVITED ADDRESS

INVITED ADDRESS

Saul Lavisky
Human Resources Research Organization

I am pleased to be here with you today, and flattered to have been invited. I am not being immodest when I alert you--beforehand--to the fact that I am not here because of any special personal expertise in the area of performance testing; I am here, rather, as the representative of an applied behavioral-science research-and-development organization which has--over the past 23+ years--developed, used, depended upon, and expanded both the theory and practice of performance testing.

The organization I represent is HumRRO--the Human Resources Research Organization. Before I get into the "meat" of my presentation, I want to say a few words about HumRRO, because I believe it will help you put my comments into perspective if you know something about my organization.

HumRRO was created in 1951 as an office of The George Washington University. Our initial mission--and our sole mission until 1967--was to conduct "human factors" research for the Department of the Army. After a few years of "covering the field," we narrowed our focus to the area of training and education because we found that this was where we could have the most immediate and most substantial impact on improving Army operations. Every officer and every enlisted member of the Army spends some time in training and/or education. In fact, when the Army is not fighting, it is training.

By the late 1950's, it was quite clear that many of the advances HumRRO was making in the psychotechnology of training and education had

relevance for civilian trainers and educators, too. (I say psychotechnology, because psychology is the basic discipline of most--but not all--members of the HumRRO professional staff.)

In 1963, I joined HumRRO in an "interpretive" role. As a long-time Army Reservist, I was familiar with the context in which HumRRO was conducting its research-and-development activities. And, with some experience in journalism, in the public schools in South Carolina, and in the National Education Association headquarters, it was presumed that I would be able to help "translate" HumRRO's work for the Army for the benefit of civilian trainers and educators.

In 1967, the HumRRO "charter" was modified to allow us to work for sponsors other than, and in addition to, the Army. And, in 1969, we separated from The George Washington University and became an independent, nonprofit R&D organization. We are headquartered in Alexandria, Virginia and we work for a variety of sponsors, military and civilian alike.

I am here today, less as an "interpreter" than as a "reporter." I want to tell you something about our experiences with performance testing, something about what we've learned over the past 23 years, and then I want to make some extrapolations from the training setting in which we've done most of our work to the education setting in which, I know, you conferees are primarily interested.

The distinction between training and education is very important, in my opinion. I want to make it now, and I will come back to it later.

Dr. Robert Glaser, in his book Training Research and Education, reminds us that the basic concern of both training and education is the modification and development of student behavior, and that both can be defined as components of "the instructional process." He suggests that

the training component refers to teaching students to perform similar or uniform behaviors. However, students display individual differences, and it is also the responsibility of instructional systems to guide the student's behavior in accordance with individual talents--in a sense, to maximize the individual differences. He refers to this activity as the educational component.

Dr. Meredith Crawford, in his chapter in the book, Psychological Principles in System Development, agrees that both training and education are concerned with human learning, and that they both share common technical problems of content and method. He makes the distinction in terms of purpose. Dr. Crawford says that training is undertaken to serve the needs of a particular system while education aims to fit persons to take their places in the many systems of society.

Both gentlemen agree that there are instructional activities which are sufficiently different from each other to warrant two different labels, despite the fact that--from a practical point of view--both the individual psychological processes involved and the technological practices to carry them out, are the same.

The key distinction for me is that the training program is intended to prepare the trainee to fit into a particular system. This makes it possible to specify the desired end-products of learning. And if you can specify these end-products, then you can design instruction to train (or "build in") the desired trainee performances, and you can design evaluation procedures to assess how well trainees can perform, and how well the training program is accomplishing its purposes. Unhappily, those of us in education do not "have it so good." We'll return to this distinction later.

Problems of terminology and definitions are not merely quibbling over "semantics." For example, let's take the term "performance tests." That's why we're here today--to talk about performance tests. As though there were any other kind. All tests are designed to elicit and/or measure performance. The original distinction was between tests that required the use of language and those that did not. The original performance test was the form-board, an intelligence test for the deaf, the language-handicapped, the foreign-born.

Even the dictionaries of psychology recognize the ambiguity of the term "performance test." One such dictionary identifies three uses of the term: (a) a test involving special apparatus, as opposed to a paper-and-pencil test; (b) a test minimizing verbal skills; (c) a work-sample test. The dictionary goes on to say that all of these uses are unfortunate because the term "performance" already means "the behavior of an examinee on a given test," and "the score of any specified examinee on a test," etc.

And yet, we here today know what modern educators mean when they use the term "performance test." Or do we? I have seen recent articles by prominent educators which dichotomized the field of achievement testing into performance tests versus "paper-and-pencil" tests, or "knowledge" tests.

As my colleague, Bill Osborn--who is with us today--points out, this kind of labeling reflects artificial distinctions and is misleading. Bill reminds us that a true performance test for many clerical tasks would also be "paper-and-pencil." And if you wanted to assess the performance of someone who operates an information center, you would have to engage in "knowledge" testing. Even a multiple-choice test can also be a performance test; take the case of a surgical assistant who has to

select the proper scalpel or other instrument at the command of the surgeon.

Incidentally, Mr. Osborn, who is the Director of the HumRRO Research Office in Louisville, Kentucky, is the HumRRO scientist who--at present--is most deeply involved in the whole area of performance testing. He will be with you throughout the day, will be one of this evening's discussants, and is much more qualified than I to answer your technical questions.

Performance testing, in the sense that I suspect most of us here think about it, has a long and honorable history. It can be traced back to ancient Greece (as so many aspects of American culture can). The medieval Guilds in Europe tested apprentices. In this country, industry has applied some form of performance testing since the Industrial Revolution. It picked up steam with the advent of the "scientific management" movement fathered by F. W. Taylor at the turn of the century. It picked up additional steam, in the military arena, during World War II, when the largest number of psychologists ever assembled on one project conducted the Army Air Forces Avaiation Psychology Program.

Throughout most of these years, the use of performance testing was pretty well restricted to occupational performances--in industry, in the military, and later in vocational education (which began as industrial-arts training). Several movements have conjoined within the past few years to bring performance testing to the forefront in general education. One of these has been the accountability movement. A second has been the behavioral-objectives movement.

Now, the notion of accountability has been around for a long time. In the traditional pattern, the school administrator has been responsible for justifying school-system performances to his political superiors--the

school board. The expectations of most Boards along these lines have been modest, and the administrators have not typically provided more justification than was required.

The newer pattern involves a more specific set of expectations, more narrowly defined, with the powers-that-be calling for meaningful indices of school-system performance. It has been expected that administrators would provide effective educational programs and would make efficient use of the resources available to them for that purpose. But now they have to prove it. And prove it not only to the school board, but to other newly-involved groups, including taxpayers, who want to know what they're getting for the additional dollars being invested in education.

It would be an understatement to report that the accountability movement has not received the wholehearted support of the educational establishment. And it should not be forgotten that the movement was not generated within the educational establishment, but was imposed on it from the outside.

Accountability is a goal-directed management process. So it is easy to see how it ties into the behavioral-objectives movement. This latter movement has had, as one of its principal purposes, making the goals of education more operational.

I use the term "operational" in the sense of "operational definition." That is, the definition specifies the operations which define the concept. In this case, the behavioral objective specifies the behavior which constitutes the objective of instruction. I know that, for this audience, I don't have to go into any detail about behaviorally-stated instructional objectives.

So, here we have the confluence of one movement which says that school administrators have to specify their purposes and accomplishments in a way that is susceptible to assessment, and another movement that says "here is the way you can specify instructional objectives to make them measurable." Don't they fit together nicely?

What has this got to do with performance testing? Well, one of the precepts of the behavioral objectives movement is that, to measure student progress, you must measure what the student can do following instruction that he could not do before. The action word there is do: what he can do following instruction that he could not do before. The objective tells us what behavior to look for, and under what conditions, and to what degree of proficiency.

Thus, the behavioral objective not only serves the instructor by making it perfectly clear what the student is supposed to accomplish, it also serves the evaluator by providing a "model" test item or set of items. Of course, in many cases, the instructor and the evaluator are one and the same person.

We'll come back to education in just a few minutes. Right now, I want to talk a little about HumRRO experience with performance testing--primarily in training, and primarily in the military training setting.

As an applied research-and-development organization, we were expected to conduct R&D that would "make a difference" in the Army's training operations. We are still in business after 23 years; we are still one of the Army's principal sources of R&D in training and education; and we are entering into contracts with an ever-increasing number of new sponsors. Those are three pretty good indices that we have, in fact, made a difference with our work.

In our early days, we spent a good bit of time working on individual Army curricula, or training programs. Essentially what we did was to apply the best of what was known about training technology to those programs of instruction that were having trouble. We would come up with a prototype, revised course; we'd compare the graduates of this experimental course with graduates from conventional courses; and if the experimental-course graduates performed better, or if they performed equally well following training which took less time or cost less money, we would recommend Army adoption and implementation.

You'll notice that I said we would compare experimental-course graduates with conventional-course graduates. This comparison is almost always made on the basis of a performance test (in the sense in which we here today are interested in that label). That is, we require the graduates of both courses to do their thing.

That "thing" is usually some facsimile of the real-world job for which the soldiers are being trained. It is a performance test in the best sense of that term. In such tests, we attempt to stimulate the information inputs to the trainee that would come to him if we were actually on the job, and to measure job output--that is, his proficiency at doing the job.

Let me try to put the test into perspective. In our course development work, we have taken what has come to be called "the systems approach." Although there are a number of variations, the HumRRO approach is shown in this paradigm.

The first step is an analysis of the operating subsystem in which the job of concern is located. This analysis provides information on the characteristics of both the hardware and human components of the system. It also gives some indication as to whether R&D efforts are

best invested in selection and classification of personnel, in human-factors engineering (that is, designing or redesigning the hardware to better fit the man), or in training. Let's assume that the answer came out "training."

In the second step, there is an analysis of the particular job about which we're concerned. We attempt to determine the inputs to the job from the rest of the system, and the outputs that are required.

It is important to note that development of the proficiency test--the performance test--is derived directly from the analysis of the job. It is in no way dependent upon what is taught in the eventual program of instruction. Ideally, we even put a separate group of researchers on this task--scientists who are not involved in the curriculum-development effort.

The final step in the paradigm is the evaluation of the new curriculum. Obviously it is not the final step in the curriculum-development activity, and there could be lines and arrows to show feedback, boxes to show revision, dissemination, and implementation. But this is the core.

There are technical problems with this kind of measurement, especially with regard to choosing the proper research design. Egon Guba has written forcefully in several AERA publications to the effect that designs that are appropriate for basic research are not appropriate for curriculum evaluation. Dr. John L. Finan addresses these problems in his chapter of the Gagne book, Psychological Principals in Systems Development. And the AERA has published seven paperback monographs on the evaluation of curriculum-development projects in which several authors also address these problems. I will not attempt to go into that kind of detail today.

A couple of years ago, I tallied up 88 training programs on which HumRRO had worked. Most of these involved the development of performance

tests--but not all. In some instaces, we produced only parts of training programs, but even in such cases, we usually went through some of the steps that we consider fundamental to the development of performance tests--that is, job analysis and task analysis activities.

We were extremely pleased, as researchers, when in 1966, the Army's training command issued a regulation on the "systems engineering of training:" that, essentially, adopted the HumRRO approach, and made it official Army doctrine. The Air Force, with which we had been sharing copies of our report, subsequently adopted a similar approach to "instructional systems development." While we can't very well take credit for the Air Force decision, we did note with some pride that more than 50 percent of the references cited in the AF Regulation were HumRRO reports.

Because the development and use of performance tests are such typical HumRRO activities, a large number of our professionals have been involved with them. However, the performance test as a subject of study in its own right has been a matter of continuing concern to Bill Osborn, Director of our Louisville Research Office. Several years ago, he chartered the major action points in the course of developing a test for training evaluation.

Let me take a moment to recap. I've explained that HumRRO began by developing and using performance tests in connection with specific training programs as part of a general overall systems approach to curriculum development. I have shown you a diagram, and will provide you with a copy of a generalized statement of the HumRRO view of what's involved in developing performance tests. I'd like to move a little closer to present-day by telling you something about a relatively new "model" for performance-based training and testing that we developed for the Army, and that is now being implemented across-the-board.

Every week, instructors in the military services are confronted with incoming classes that must be taught a considerable amount in short and relatively fixed periods of time. These classes are usually quite heterogeneous with respect to students' educational background and learning aptitude. From earlier research, by a large number of individuals and organizations, it was apparent that the traditional military lock-step, lecture-demonstrate-practice-test approach to instruction would not be particularly effective for trainees at either end of the ability spectrum--the low-aptitude and high-aptitude personnel.

HumRRO was asked to come up with a new approach to Army training. It had to be both effective and efficient. And there were other constraints. It couldn't cost any more than current instruction. It couldn't require instructors of higher caliber or greater sophistication in training. It could not require any significant increase in the amount of operational equipment for practice, nor could it require any extension of the training period, or expensive instructional hardware or software. In sum, the new approach had to be fashioned out of the currently available resources.

Under such constraints, the new approach, or "model," as we like to call it, evolved as one in which the instruction of trainees by other trainees is a central feature--that is, peer instruction.

There are six principal features to the model, in addition to peer instruction:

(1) Modular Sequencing. The course is organized around a series of job-performance stations that represent the various duties performed by a person competent in the job. The number of stations is determined by the number of coherent sub-jobs in a specialty. Since each station represents discrete sets of tasks, a trainee can enter the system at any point.

(2) Self-Pacing. The period of time a trainee spends at any station depends on how long it takes him to learn to perform the tasks.

(3) Insistence on Mastery. Each trainee undergoes a proficiency test (that is to say, a performance test) when he is satisfied that he has learned a task. He must demonstrate that he has mastered the necessary skills before he is allowed to proceed to the next task in the sequence. If he fails any test, he must review and practice until he can pass. Incidentally, for quality-control purposes, the tests are administered not by the peer instructors, but by full-time cadre members, who are on hand as training supervisors.

(4) Rapid and Detailed Feedback to Trainees. Since proficiency tests follow each task, the trainee knows immediately whether he has learned the required skills.

(5) Rapid and Detailed Feedback to Instructors. Since the trainer/supervisor administers the proficiency test, he knows immediately whether the instruction has been successful.

(6) Functional Context Training. Job-performance stations represent actual on-the-job duties that must be performed, so the trainee actually learns the required skills and knowledges in a job-like setting.

This new model was field-tested at Fort Ord, California, with soldiers training to be Field Wiremen. It produced graduates who were markedly more competent at their job than conventionally trained wiremen. At the same time, it also reduced training time, training costs, academic recycling, and academic failures.

The Army immediately adopted the prototype program for all its field-wiremen training, throughout the United States. It also directed that this new, performance-oriented approach to training and testing be adopted throughout the Army. In recent months, we have been helping Army

training managers implement this new model in courses for cooks, mechanics, heavy-equipment operators, air defense technicians, and infantry.

I am sure you can recognize the key role that performance testing plays in this approach to training.

In 1971-72, we moved the model outside the Army and tested it in the public schools in a course on the office cluster of business occupations. This test was conducted in the Pacific Grove Unified School District. Test results indicated that this performance-based instructional model produced graduates with statistically significantly superior job knowledge, who were dramatically superior in job performance than their conventionally trained peers.

We have introduced this model into a junior college in Vermont where the emphasis is on Adult Basic Education, and on occupational/vocational education, primarily for rural white adults. We have also introduced this model into a Community Action Project in Alabama, where the concern is with training women for occupations as household workers.

I recognize that your interest today is in performance tests, per se, rather than their role in programs of this kind, no matter how innovative or effective. But my point here is, simply, that it is performance testing that is driving the system.

To this point, I have covered the past and the present. What about the future?

I come back again to Bill Osborn, and one of his current projects. In fact, I've quoted him and borrowed from him so often that it is clear to me now that it is he, rather than I, who should be addressing you. However, having typed this much manuscript with two fingers and a thumb, I refuse to relinquish the podium. I will press on.

Bill points out that the logic of developing a performance test is simple. You conduct a job/task analysis, recreate the job task in a test setting, ask the trainee to perform the task, then record whether he did it or not. Unhappily there are many, many reasons why performance testing is not that simple.

Let me cite but one example--and a "simple" one, at that--teaching someone to drive an automobile. Dr. A. James McKnight, on a 1971 HumRRO project, conducted a comprehensive analysis of the driver's task in order to identify critical driving behaviors from which instructional objectives and test items could be drived. He and his colleagues found that in simply driving on an open highway there are more than 1,700 specific driving behaviors. You can imagine the number of instructional objectives and the number of test items that would have been required if some process of distillation, some determination of criticality, had not been undertaken. And there is little in the performance-testing literature or job-analysis literature to guide the scientist in this distillation process.

The two major evaluation tools the instructor has available are job-knowledge tests and job-performance tests. There is a question as to how results from the two types of tests correlate. Some researchers, in some settings, have found correlations so low as to indicate that job-knowledge tests are practically worthless for assessing individual proficiency. Other researchers, in other settings, have found the correlation reasonably high.

Practically everyone agrees that a performance test is, in some way, <u>better</u> than a knowledge test. I think we would all feel happier, as instructors, if we could have our students <u>do</u> the job rather than <u>tell us about</u> doing the job. But the typical performance test is more

expensive than the knowledge test. It sometimes requires too much equipment, too many test administrators. And sometimes. the level of professional skill needed to develop and supervise administration of performance tests is simply not readily available.

As Osborn points out, the training manager is faced with a choice between a practical evaluation tool with questionable validity on the one hand (the knowledge test), and an impractical tool with high validity on the other hand (the performance test). He feels that this Hobson's choice presents a false dilemma--that there are other solutions that lie in between these two extremes. He has proposed the concept of the "synthetic" test and is busily engaged these days in testing the concept.

In fact, Bill suggests that there are a number of alternatives which fit between the two extremes, each one combining a differing mixture of validity and feasibility. Let me give an example.

Let's assume we have the following performance objective. "Given binoculars, paper and pencil, and 20 targets in various degrees of concealment and orientation, at ranges of 500 to 2500 meters, the soldier will estimate and report the range to each target, accurate within two meters on 16 targets within 10 minutes."

This objective consists of eight behaviors. If we took the soldier onto the range and conducted a full field test, we would be able to assess his performance on all eight. However, given a large number of soldiers to be tested, and only limited resources, the typical reaction is to conduct a paper-and-pencil test of the soldier's understanding of the mil relation formula. This test addresses only two of the eight component behaviors, but it is easy to administer.

I've already talked about the method 1 (the field test) and method 6 (the paper-and-pencil test). The in-between methods represent

alternatives of intermediate complexity. They were fabricated by considering economically available ways of eliciting each skill and knowledge, and then synthesizing them into a test method.

In weighing these alternatives, note that as the simplicity of the test method increases, information on some component behaviors is lost. The simpler we try to get, the more information we lose. We eventually reach a point where it doesn't even make sense to give a test. Also, the simpler the test method, the more diagnostic information we lose--that is, information that could help us identify where our training program needs improvement.

The concept seems reasonable. Mr. Osborn is doing more in the way of conceptual development, is seeking empirical verification of his notions, and will eventually codify procedures under which test developers can use "synthetic" performance tests.

I have taken longer than I intended to reach this point, but before I conclude, I want to return briefly to the distinction between training and education that I made earlier. You remember that both Dr. Glaser and Dr. Crawford made the point that it was easier to identify training requirements than educational ones. The job-analysis/task-analysis approach doesn't have much application in the general-education, liberal-arts fields--at least not yet, so far as I can see.

If you were to view instruction as taking place somewhere along a continuum that runs from the specificity of training to the generality of education, you would find the concept of performance testing increasingly difficult to apply as you move from training toward education. It shouldn't be necessary for me to remind any of you that a test--even a performance test--is only one tool for evaluation. This is even more true when you are appraising individuals instead of instructional programs.

Performance testing is only a leaf on the twig of tests and measurement, on the branch of evaluation, on the tree of instruction. And, if I may be forgiven another simile, I hope that none of us will even be accused of being like the small boy who is given a hammer as a present; it's amazing how many things he can find around the house that need a good pounding.

The examples I've cited for you today have all come from the training end of that training-education continuum I mentioned. To move toward the education end will take time, effort, imagination, and ingenuity. But those of us in the performance testing business won't have to go it alone. We're in good company because the instructional technologists and the systems analysts are all wrestling with the same problem. If, and when, we and they develop tools and techniques for reducing our global educational goals to discrete, behavioral objectives, the rest of the job will be much, much easier.

In the original charge given me by the Clearinghouse, I was asked to conclude my presentation by identifying major gaps in our understanding of performance testing, and to suggest directions that future R&D might take. I have made several stabs in that direction, but must confess that I am unable to carry out that assignment. I can only identify gaps that strike me as important. I suspect that, given our multiple purposes for wanting to use performance tests, we might each develop a different list. However, for what it's worth, here's my list.

First, I would like to see someone undertake a state-of-the-art survey of performance testing, and come up with a handbook or how-to-do-it manual that evaluators and teachers could use today. Not all evaluators are as sophisticated as they should be, and not all teachers have the time to delve deeply into the subject.

Second, we need that missing third taxonomy. We have Bloom, et al., on the cognitive domain. And we have Krathwohl, et al., on the affective domain. But we don't yet have a suitable taxonomy for the psychomotor domain. Dr. Fleischman and his colleagues at the American Institute of Research have been working in this area for some years, and their reports are both interesting and useful. They may be on the verge of the kind of taxonomy I'm talking about but, in any event, we need it, and soon.

Third, it seems to me that the major problem faced by those who want to use performance tests in education is the problem of criterion. By your interest in performance testing, you have indicated an interest in moving education from norm-referenced tests to criterion-referenced tests. But to do this, you must have a criterion. And if your efforts are to be fruitful, you must have an appropriately relevant criterion.

Our colleagues in the human-factors engineering field are interested primarily in human performance in man-machine systems. This is only one of the kinds of human performance in which we, as educators, are interested. And yet, in their relatively small area, they have considerable difficulty finding appropriate criteria on which to validate their proficiency and predictive tests. How much more difficult our job--we who take all of education as our territory.

Fourth, I come to a closely-related problem area (one I've already mentioned): the specification of observable, measurable instructional objectives. We must find a way to operationalize our beautiful-but-abstract educational goals. There is an element of truth in the accusations that we have, thus far, been able to develop measurable objectives only for the "trivial" outcomes of education. There are, in fact, important outcomes that we have not yet been able to express in behavioral terms.

And the more abstract the outcome, the more difficult the task--both for instructional design, instruction, and evaluation.

Fifth, the two foregoing problem areas can be incorporated along with the problem of determining what ought to be the goals of American education. This is not a problem for which educational evaluators have any unique responsibility. On the contrary, everyone in education (and outside it, too) has some degree of responsibility for determining the most appropriate goals for American education. But we have some unique tools and several potentially useful methodologies to offer. And we are reasonably committed to the "scientific approach" which, I feel, is badly needed to leaven the mixture of arm-chair philosophy, common sense, and vested interests with which this topic is commonly addressed.

One recurring suggestion has been that we concentrate on competency in adult life. David McClelland spoke to this point in his 1973 American Psychologist article. This is not a novel suggestion. Between 1915 and 1919, the NEA Committee of the Economy of Time sought to identify what adults do, and what they need to know, and to use this information in establishing goals for American education. In my own mind, I date the beginning of scientific curriculum-making by the work of that Committee. I commend its four-volume report and the McClelland article to your attention.

In conclusion, let me quote from Dr. Earl Alluisi, formerly of the University of Louisville, and now of the University of Virginia. In a 1967 article in the journal, Human Factors, he said:

"'Performance assessment' is one of the most important and difficult areas of current research. It is important in its own right, as any supervisor who has been called upon to justify the ratings of his workers can attest. It is important also because it is the crux of the 'criterion

problem' for so much other work; the final validation of selection and training techniques depends upon the assessment of the performance of men who have been differently selected and trained. The final validation of an improved, human-engineered, man-machine system depends upon it . . . The assessment of man's behavior in the meaningful performance of complex tasks has challenged physiologists, engineers, and psychologists for many years. The task has been recognized as a difficult one; the problems have been formidable; and the solutions have been ephemeral . . . Considerable quantities of good and respectable research have been published . . . (which) advanced science generally, but it has failed to provide any significant progress towards performance assessment . . ."

If we come away from this conference having advanced the ball only a matter of inches, it will have been worthwhile.

SMALL GROUP DISCUSSION REPORTS ON PROBLEMS, ISSUES AND NEEDED RESEARCH AND DEVELOPMENT IN APPLIED PERFORMANCE TESTING

Administrator Group: Gerald H. Lunney, Reporter

Like so many things, it's so easy to say yes to being a small group reporter but it is so hard to do. We had, I think, a very stimulating session. I'm just going to run through some random thoughts.

Interestingly enough, for a group of administrators, we spent a good deal of time on two issues. The first one was cost. References were made to cost twenty times in our group. There were some interesting concerns relative to costs and to the whole question of APT. Cost is involved because when you measure behavior, you need people to conduct the measurement, and often they are not trained to adequately observe what is going on. One general issue that was raised was the problem of reliability of graders. Along with that and other concerns that affect cost is the question of what criteria relate to good performance. How are you going to clarify good performance so that everybody knows what it is and when it has taken place.

The other major issue for administrators was politics. Part of this concern came from the fact that a great deal of the interest expressed in APT has come, as we have mentioned before, from external agents. We got into another topic which I have noted here as "standardization plus and standardization minus." Standardization plus was concerned with the fact that if you had different people defining appropriate performance, how do you arrive at an acceptable definition? How are we going to be able to say that if a student passes the test in City A and then moves to City B, that his performance is acceptable?

Standardization minus, regarding the actual behaviors of people, raises personally or culturally bound questions about performance. Since we have a diversity in cultures and we are talking about whether we can really establish some overall standards, how can we say that once a student has passed the tests he can perform adequately in different kinds of settings, interacting with different kinds of people?

Another topic we discussed was who should set the standards of behavior. How should they be established? For example, who is going to establish the standards for mathematics? Is it going to be the parents with a wide variety of needs or desires for their children? Is it going to be the school? Is it going to be those people who will ultimately employ the students? We can't differ from the curriculum in applied performance testing. As we develop the curriculum in conjunction with applied performance testing we are talking about a single package and not different things. Someone raised a point about freezing the appropriate behavior in time--the question of whether current adult behavior is a sufficient standard for judging performance.

The last point relates to the appropriate starting point: Are we doing the same thing with applied performance testing that we did with criterion-referenced testing? Are we trying to legitimize it by making sure it fits everything that we learned when we took the first three courses in tests and measurements? Or, is it sufficiently different that we should perhaps hold back a little and make sure we have appropriate testing gear. Can APT stand by itself and can't we legitimize it on that basis--and not on the basis of how well we can fit it to what we learned way back when.

Curriculum Group: Craig Gjerde, Reporter

Since we were called a curriculum group we tended to get away from the testing and get into the planning of tests. We had an interesting, but hurried discussion of many points which have already been discussed. One big question was, How do you define skills needed by the student, especially in regard to adult life or the many styles of life those people might live as adults? We suggest that perhaps those life skills should not be defined by people in the traditional teaching disciplines. They should somehow be linked to survival education; we didn't define explicitly what we mean by "applied" performance testing, but I think we felt that the term referred more to the long-term survival value of the education than to the current academic emphasis.

Another question that the administrators should perhaps have considered was, How do you handle differences in completion times that could occur if you get into applied performance testing? Do you allow students to leave high school when they are thirteen years old? There was also some discussion about how long it might take to install applied performance testing in our educational system. Some people thought that it might take us a long time to rethink our educational values and develop appropriate applied performance testing strategies.

One basic issue that comes up over and over is, Who are the experts that are going to develop this applied curriculum? How are we going to somehow agree on what these survival value skills are? We have to recognize that there will be many political and social pressures that will resist the change.

We identified some areas in which research and development efforts are needed: defining the kind of staff and facilities needed to implement applied performance testing, and developing instructional modules based on the several different approaches to organizing materials.

Measurement and Evaluation Group I: Sarah S. Knight, Reporter

Being more or less measurement oriented we started with the evaluation guidelines for applied performance testing. The first thing we concluded was that the outline form might not be very helpful. It seemed overly complex; the wording could be simplified. We talked about the audience for these guidelines, and determined that they were written for the technician in tests and measurements. It was suggested that we might want to modify the definition of applied performance testing, so we were not restricted to either simulated or real situations but could allow for inferential testing.

We discussed test content with respect to minority groups, and there was some concern regarding possible over-emphasis on minority group content, which might actually restrict the kinds of things we could test for and result in a test that could only be locally applied. It was suggested that we take a look at the EEOC guidelines (1970) for employee selection procedures.

Then we talked about problems. One problem concerned developing performance tests. Probably one of the more functional ways to develop such tests is to concentrate on the parts of a task that a person does incorrectly (i.e., to concentrate on the errors rather than the total content of the task). Also, the question was raised whether we should concentrate on process or product in terms of applied performance. We concluded, as the curriculum people did, that reliability was a crucial problem.

In talking about applications in public schools, we got into a long discussion about what constituted basic skills and how we were going to test them.

Measurement and Evaluation Group II: Richard L. Stiles, Reporter

Essentially, we discussed two questions: What do we test? and Do the tests have content validity? Under these headings we dealt with competency-based instruction and prerequisite performance skills.

Also, we discussed who should set priorities in identifying important performance, and for what purpose. Should target performance be set at the school level, district level, county level, state level, or national level? Who is buying the can of dog food--the dog or the owner?

What should we test? I think testing what the learner will be predisposed to perform is important. That is, test what he will do in the future versus what he can do now. I think in terms of accountability, we can say that the schools be responsible for current performance. We can't guarantee what learners will do when they go out into society.

There is a problem in defining literacy. For example, do you need inferential skills to be considered functionally literate? Again, in applying applied performance testing in public schools at lower educational levels, if you have trouble deciding at the high level what constitutes basic skills, you will not be able to identify prerequisites. We are going to have to do a little backtracking.

With respect to issues in public schools, to what extent are the public schools responsible for those identified skills? That is, many performances might not actually be a part of what public schools now believe they should be doing.

In terms of research efforts, it was suggested that we need more emphasis on research regarding naturalistic observation. We need to consider doing longitudinal studies on performance testing. We need to identify situations in which applied performance testing is appropriate.

V.

DISCUSSIONS OF PROBLEMS, ISSUES, AND NEEDED RESEARCH AND DEVELOPMENT IN APPLIED PERFORMANCE TESTING

Joseph L. Boyd, Jr.
Educational Testing Service

One of my major concerns regarding the development and use of performance tests, about which I had intended to speak tonight, has been addressed in the paper "Criterion Guidelines for Evaluation of Applied Performance Test Materials and Procedures." In presenting the paper for discussion, CAPT has taken a very important step toward increasing the development and use of performance tests in schools. Congratulations to CAPT--even though they stole my thunder!

In emphasizing the development of testing instruments involving real-life simulations we must not lose sight of the fact that some paper and pencil tests can require complex performance and are, in that sense, a kind of "performance" test. Examples include the "patient management" problems of the National Board of Medical Examiners tests,[1] and several nursing specialty certification programs. These tests are modern variants of the "tab" test, whereby the examinee makes a response choice and obtains additional information. These tests are variously referred to as programmed tests, or variable sequence tests. The ultimate failure on such a test occurs when a physician erases his final choice in a series of choices and gets the additional information: "patient expired!"

This type of test has also been developed to assess other kinds of diagnostic skills. British radio repairmen take a programmed test as part of the procedure for occupational licensing. The number of examinees had made the trouble-shooting test with real radios an unmanageable

[1]Hubbard, John P., "Programmed Testing in the Examinations of the National Board of Medical Examiners." Proceedings of the 1963 Invitational Conference of Testing Problems.

task. Motor vehicles, guided missile electronic and hydraulic systems, radar, and television fault diagnosis tests are also being used. On a small scale, programmed testing can be done with a response on one side of a card, and the added information on the other.

I would like to make another observation--regarding standardization of tests. I speak not of the statistical treatment of test results, but of specifying to the examinee exactly what he or she is to do, and observing and grading the performance, or product, or both in a systematic, predetermined manner. To me, this is performance testing. I recently reviewed a paper in which the author failed to differentiate between standardized testing such as I have defined it, and observation of unspecified, undirected student behavior. The latter activity could never be construed as "performance testing," as I see it.

I appreciate the opportunity I've had today to hear and be heard. CAPT and NCME have done a great service to education in hosting this meeting. I am taking away from this meeting much more than I brought. Thank you.

Hulda Grobman[1]
University of Illinois College of Medicine

Aspects of performance tests that may be noteworthy, in addition to those mentioned by Mr. Lavisky:

1. It is far more difficult to standardize test administration conditions for performance tests than for conventional tests insofar as the critical (significant) variables are concerned. What is supposed to happen in the way of environment and process, what is reported to have happened, and what actually did happen may not be entirely congruent. Non-events--things that are supposed to happen but did not--may be frequent. Feedback from examinees concerning test administration conditions may be one way of checking on gross omissions or commissions.

2. What is a satisfactory correlation between test performance and actual on-the-job performance? A correlation appropriate for one purpose may not be appropriate for another. Thus, a correlation of .65 or .70 between performance test and on-the-job performance for a job that is relatively closely supervised or non-critical (in terms of cost of error--material and human) may be appropriate. But a similar correlation may be inappropriate in an area where error is critical (surgery, or navigation of a plane). Also, it should be kept in mind that correlation does not imply a cause-effect relationship. A correlation between a performance test and on-the-job performance may reflect a third variable which may not always be present, and so the correlation may change unexpectedly.

---

[1]Presently Professor of Health Education, St. Louis University Medical Center

3. The selection of tasks for performance tests (the portion of the universe to be selected), and the repitition of tasks present problems. Reliability in the test-retest context presents problems since human performance is not necessarily reliable. People have good days and bad days. How many times should a task or kind of task be repeated to validly assess its mastery? And tasks within a job may be unrelated in terms of mastery, so that internal consistency measures may be inappropriate.

4. Though content validity might appear to be self-evident for performance tests, such validity may not exist. The tasks to be performed may not, in fact, be a necessary component or standard.

5. Scoring mechanisms for performance tests require more systematic concern than may be self-evident. Scoring is probably a more complex concern than is the case with conventional tests. In addition to the question of obtaining reliable scoring, is the question of whether the examinee should be permitted to continue a test after committing a serious error at some point before completion. Allowing him to continue may waste resources and endanger the examinee or a subject he is interacting with. What are go/no-go points? What are valid criteria for establishing these? How should elements of the exam be weighted? Are all equal? Are some absolute requisites and other desirable non-requisites? If passing is based on total score, we may pass a student who ruins his machine or kills his patient.

6. As in more conventional testing, the performance test may require modification to reflect whether it is for diagnostic and/or certifying purposes. For certifying purposes, a product may be all that is needed to judge adequacy of performance; for diagnostic purposes, the product alone may provide sufficient data.

7. Because the format of performance tests will be a new experience for many examinees, there should be prior explanation of the format and, if at all possible, a practice dry run using the performance format. Without such an advance tryout, the test may be one of the examinee's adaptability or testwiseness rather than of his ability to carry out a specified job.

8. Even teachers who have been observing performance for many years may not be accurate observers. Observing is a learned skill; it requires practice and uniform criteria. Such uniformity of interpretation, whether of product or process, cannot be assumed. Performance tests lend themselves to the halo effect at least as readily as essay tests, the grading of which is notoriously unreliable. Like essay tests, performance tests seem simple to construct, and this may be the case. However, the scale for judging performance is not. And neither is it simple to achieve congruence among raters or for one rater over time. Without such congruence, the test is invalid.

9. Performance testing is admittedly expensive, far more so than conventional paper-and-pencil testing, in terms of the facilities and the time required of examinee and examiner. However, failure to use performance tests may be still more expensive--though the cost may not be as obvious.

Areas that might be explored by CAPT in the coming year:

1. Some refinement and elaboration of the Guidelines for the Evaluation of Applied Performance Test Materials and Procedures is needed. It is hoped that a document would be produced comparable to the APA Standards for Educational and Psychological Tests--keeping firmly in mind the differences between performance tests and conventional tests, and the fact that many performance tests are criterion-referenced rather than

norm-referenced. Thus, the Guidelines should not be bound by conventional test practices and standards where these are inappropriate to the purposes or format of performance tests.

2. Preparation of a basic text for the classroom teacher on how to write--and how not to write--performance tests is desirable. An annotated bibliography is not sufficient, since much of what has been written to date about performance testing is buried in materials concerning conventional testing. And there is probably no existing how-to-do-it source to prepare a performance test writer appropriately or efficiently. A second, more sophisticated and detailed text for the test specialist, covering preparation, use, and interpretation of results of performance testing is also needed.

3. It would be useful to have a section of the CAPT Newsletter or a comparable medium devoted to exchange ideas about performance testing; a publication similar to the UCLA Evaluation Comments might be appropriate.

4. Some investigation should be undertaken concerning various legal aspects of performance testing. Two aspects requiring early attention come to mind:

The first concerns records of performance. In conventional testing, if there is some question regarding the accuracy of test scoring, one can return to the answer sheets to verify the scoring; for certifying tests, it is conventional practice to retain answer sheets for some time in the event that questions concerning accuracy of scoring arise. For performance tests in which the test is of process, the only record is the examiner's recording sheet. If his accuracy or objectivity is questioned, will additional evidence be needed to verify or justify the score? For certifying examinations this could become a critical

issue. Second, an important aspect of performance is how an individual operates in crisis situations. However, to subject an examinee to performance tests simulating severe stress may, however realistic, be inappropriate and highly unacceptable. To what extent can/should crises be incorporated into testing? And what options are more acceptable?

5. Some types of performance tests need not be kept secure. For example, a checklist for a given job or product in effect provides the objectives and a learning resource as well as the rating system for the performance test. However, for problem-solving skills, if the problem is to be a new one to the examinee, so that he can demonstrate problem-solving ability, the test must remain secure.

It would be useful, given the expense of developing tests, to have a mechanism for sharing secure tests while still maintaining security. The design and implementation of such a system would be a major contribution to the field of performance testing.

Dr. Ruth Nickse
Syracuse University

I feel if this meeting had been held a year ago I would not have ventured forth where angels fear to tread. I had never heard of applied performance testing, and in fact, I didn't know that was what I was doing until I received the material from CAPT not too long ago. I will tell you what my charge was and what I did and then you can throw your rocks and stones because we certainly have, in our effort to do something different in testing, thrown out the baby with the bath water.

When I joined the group in Syracuse my charge was to design an assessment system that would provide an opportunity for adults to demonstrate what they knew and could do regardless of where they had learned it, in certain required areas such as computation, communication and life skills; and secondly, to grant a regular high school diploma to ratify this learning. What we did was to develop a new kind of testing program based on some assumptions we had about adult learners. Our program is full of assumptions. We feel very comfortable with them right now, but of course they are very questionable. We assumed that adult learners were test anxious having come through the American school system; that they were rebellious because they had been subjected to GED exams when they were perfectly fine auto mechanics; and some did not care about--or get the proper answer about--the amount of wood pulp in the State of Oregon in 1922. We figured that they were busy with full-time jobs and families and had little time to sit around testing rooms. We figured that they were highly motivated to work for a high school diploma after many years out of school. We figured that because they were adults

they could be responsible for their learning and testing situation. And, we figured that they needed an opportunity to choose assessment molds which would best enable them to present their skills and competencies. Above all, and this was our biggest assumption--we felt that they were competent in life skills by virtue of having lived and worked in the community.

If you start with these assumptions, you are free to design what we call an open assessment system. But you have to go along with all of these assumptions. We decided that we didn't want to created another GED. Many of the persons we hoped to reach as a target population had had sad experiences with the GED and other kinds of standardized tests. Since we were free to dream wild and big, we did. Our objectives were to design an assessment system responsive to adult learners, to give learners some control over the testing environment, to make the assessment process a learning experience, to relate assessment form and content to the concerns of adults and to make the testing process humane insofar as we could do it.

In order to reach the objective of giving learners control over the testing environment, we designed diagnostic and final assessment instruments and processes that are initiated on demand and are self-paced. In our system, learners have several assessment options. In order to make the assessment process a learning process we have told the learners in advance the 64 competencies that they will be required to demonstrate. Throughout the program we keep them thoroughly informed of their progress in demonstrating the 64 competencies.

Some of the tasks that we all face as adults are changing residence, finding a place to live, finding a job, developing consumer awareness and maintaining personal health. We used simulations and we used oral

interviews as part of our procedures. In order to make the testing process humane, we went to individualized testing. We helped the learners assume responsibility for their own progress because they initiated the request for testing. We give them continuous feedback and success experiences because we hope to design this on their strengths and not on their weaknesses. Too often, I think, testing makes it easier for us test designers to work on error; it doesn't make the person being tested feel so good.

The distinctive features of our external diploma assessment process are these; we always talk about the good conditions for testing. Well, the temperature has to be right, the light has to be right, the distraction level has to be down, and so on. The best place for that is at home, so we designed flexibility in time and location of testing by allowing the adults the opportunity to establish some of the conditions of testing by taking three of the tests that purport to assess the 64 competencies at home. They can take the tests in our office, but if it is more comfortable at home, they can do it at home. After all, if you are working on two jobs, the time you have for testing is pretty short. Two of the tests are oral, because some people do not do well when they have to write, but they do speak well. If it is a matter of health or related health competencies for yourself and family it is just as valid, I think, to discuss these kinds of things as to write them down or to choose the correct multiple choice answer.

We felt it was important to have open information on the requirements so the competencies would be explicit and open to discussion. Learners are given a copy of these competencies to take home. As a matter of fact, as one of the diagnostic instruments we have a self-rating checklist. It is amazing to find out that adults realize what they don't

know and are willing to mark that down as a weak area as long as there is no penalty attached.

We have flexibility in scoring. There are some right-answer questions in math. Of course, you have to be 100 percent right. It's like in the old days when mother said that you had to eat everything on your plate. In our system math is one of those things. You must demonstrate it and you must demonstrate it 100 percent. Some of the answers to our questions are merely documentation. If you wish to ask the question, "What is evidence of having participated in the community as a responsible voter?", one of the things that you might document is whether the person has a Voter's registration card.

We have continuous feedback after each of the take-home tasks. There is a spotcheck in which some of the most vital competencies are tested in the office again. We know that the wife could give a little help to the husband who is studying for his external diploma, so we do ask them to demonstrate some of the critical competencies back at the office. They don't object to that, as a matter of fact.

We offer, of course, the first competency-based diplomas in the country. We have scooped Oregon which is going to be giving diplomas in 1978 for demonstrating life skills competencies. We feel this is an important direction for adult education; the implications for secondary and elementary school curriculum I will leave to those persons who are involved in this at the state level. However, I want to draw your attention to the work of Dr. Norvelle Northcutt in Austin, Texas, who has partly answered the question about what adults need to know to function in our society. The results of his study, which identifies some 75 competencies that adults probably need to function successfully, will be out in December of this year. His national survey of adult competencies will

be shocking to some and not surprising at all to others. But, if you are concerned about validation of competencies, his work offers a valuable resource for you.

In our work we have been confronted with several kinds of problems and we request your help on these. I don't think any of them are new. One of the problems is that we do not have a process for good task analysis. Since competencies are not leveled in our system and since adults need to read passages on many different levels--from road signs to the domain of leases--we need a process for good task analysis.

Marilyn Lichtman's reading test is probably the first one that I have seen that confronts the nine or ten different domains of reading in which adults must achieve in order to be successful in their daily lives. It has been the most useful standardized test booklet that we can find. If you are interested in such a test, you should look it up. It is a self-paced, self-initiated test which adults, in my experience, are pleased to take because they find it relevant to their needs and interests. But, of course, it doesn't break down reading tasks into small prerequisite skills. We do need a method of breaking down competencies into prerequisite skills; that is an enormous job.

We need some criterion samples. We need a behavior analysis of what constitutes good or poor performance. And then we have to ask ourselves, "Why are we asking ourselves that?" I think there is a tremendous amount of value judgment in education, notably in the field of applied performance testing. Our 64 competencies were selected by a task force of persons who probably came from similar backgrounds and valued the same competencies. Whether those competencies are truly representative of what all people in our adult society need to know I'm not sure. We should remember that we make value judgments each time we select a

group of competencies and then decide to legislate them, or by fiat, label them a curriculum.

We must empirically evaluate what people really do on their jobs and in their lives. What one thinks they do and what they actually do may be different. Testing should correspond to reality. Some behaviors are probably common across certain occupational fields, but until we do empirically validate those things I think we have to hold ourselves in check and realize that we are making value judgments all the time.

We really do need to review competencies at frequent intervals. If we are to say that they represent what adults in this society need to know, I think they might need changing every month--or at least every year or two. Those of us involved in the explication and testing of competencies must realize that we are in for the long haul and budget some money for regular reviewing. I think it is very important that we have good behavioral objectives with precise criteria defined. No matter how we do it we can't afford that step, because as Bill Osborn said, "out of the objective comes the testing." But the questions behind that are, Why this particular item? Why that competency? Who values it, and who sets up the criteria? Those are big questions.

I won't add any more to my plea except that I see applied performance testing as a chance to humanize assessment. Of course, that works in exactly the opposite way from cost accounting and our concern with group tests. My concern has been with the adults who take the test. The beautiful test is a wonderful thing and I value it too, but the persons taking the test are equally valuable and their needs as test takers should be considered.

I picked up a quotation relevant to my work from someone who has written a little book on how to conduct oral interviews. "We cannot

humanize assessment without taking risks of abuse. The problem is to preserve humanity while enhancing validity." If CAPT can tell me how to do that, then it has been organized for a good reason.

William C. Osborn
Human Resources Research Organization

As one who has been involved for several years in the development of performance tests--chiefly in connection with Army training evaluation--I see the problems of performance-based measurement in the field of educational evaluation to be essentially unchanged during that time. Most are practical problems encountered in trying to provide what might be termed efficient tests--that is, tests which are valid and reliable, but also usable in the sense of evaluating the proficiency of large numbers of people at minimum cost in time and resources. Achieving a balance between test quality and administration economy lies at the heart of the performance testing problem.

Although performance tests have other purposes, they are used chiefly in evaluating training and educational outcomes. Following training on a job- or life-task, a student is normally required to demonstrate proficiency on that task before being advanced to the next stage of learning, or ultimately, out of school and into the world of work. The development and use of such tests would seem to be straightforward: the job- or life-relevant conditions for task performance are specified and an acceptable criterion of performance defined. The student's performance is then evaluated according to the established criterion. Unfortunately, the nature of certain job- or life-tasks, together with time and cost constraints, often create problems for the test developer. In circumventing these problems he may resort to simplistic test procedures of questionable reliability or validity. The seriousness of this problem is reflected in the fact that such comprises very frequently

occur--apparently either because of inadequate regard for the price one pays in diminishing reliability and validity, or because developers are not aware of alternate approaches.

This evening I would like to summarize briefly four aspects of performance test development that I consider essential to the practical achievement of reliable and valid measures. Please bear in mind that my observations will be limited to test development for individual tasks and will not touch on other aspects of reliability and validity--such as sampling of the job task domain or replications of test performance--which pertain to testing on an aggregate of tasks or an entire job.

Test Method

The first critical aspect of a performance test to be considered pertains to the directness or relevance of what I will call the method of testing. A test method is relevant or direct if it requires performance identical to that specified in the actual job- or life-task. The scope and fidelity of actual job or life conditions presented and the realism of the response medium used determine the directness of the testing method.

In a training or other performance assessment setting, limited resources often prevent a direct task enactment method of testing. Indirect methods, involving partial task performance or simulation of task conditions are often used. Such methods commonly measure performance only on the more testable part of the task. Paper-and-pencil knowledge tests on tasks requiring both knowledge and skill represent the most flagrant example of indirect testing. Tests of job knowledge are relatively inexpensive and have exceptional psychometric properties. Yet, for obvious reasons, we would never consider licensing a man to fly a

plane or drive a car merely on the basis of a knowledge test. But why then, in other job or job task areas, do we tend to accept knowledge as a valid measure of performance capability? The chief reason is cost. A performance test presents the real work environment with all its cues, then elicits actual job behavior as directly as possible. But representation of the real world is expensive. Educational and personnel administrators tend to think performance tests require too much in the way of equipment, personnel and time to justify their use. To insist, however, that a test of job knowledge is the only alternative reflects a false dilemma.

For any given job task several alternative testing methods are available. These will run the gamut from an expensive but fully relevant performance test to a relatively inexpensive but marginally valid knowledge test. Elsewhere, I have described an approach to devising alternate test methods, based on the concepts of simulation and task-element sampling. I have collectively termed such measures Synthetic Performance Tests. The intention is to connote a process of synthesis by which the substructure of a job task becomes the basis for selectively constructing alternative forms of a test, each representing (at least theoretically) a more or less optimal blend of validity and feasibility. In some cases this optimal blend may be achieved through simulations; that is, by substitution stimuli in either the task display or the surround, or by requiring a substitute response. In other cases, performance may be efficiently measured by testing on a subset of task elements, regardless of whether simulation is used. Thus, synthetically generated alternatives to fully relevant performance tests may vary in two major dimensions: fidelity and scope.

Consider, for example, an electronic troubleshooting task. Knowing the correct test sequence for isolating a faulty equipment component is only part of the task. Among other task elements the troubleshooter must also be able to place the test-set in operation, establish a good connection at the test points, and correctly interpret the test readouts. Can this type of job task be adequately--that is, validly--tested with a traditional verbally formatted test of job knowledge? I would say no. In fact, experience may reveal that, on the job, a frequent case of faulty troubleshooting is the inability of the troubleshooter to establish good connections at the test points--an essentially physical or manipulative element in the task performance. So, assuming the test developer cannot afford the luxury of a direct, hands-on method of testing, the important thing is that he does not immediately revert to the typical knowledge test. He should use his inventiveness in devising alternative testing methods that call for demonstrated behavior as similar as possible to that required in task performance. Pictorial, graphic, or even low cost three dimensional simulators should be considered. The developer may assess the relevance of these synthetic options by checking the breadth and criticality of task elements measured by a particular method.

Only in this way, it seems to me, can test developers arrive at economical methods of proficiency testing while maintaining an acceptable level of content validity.

Test Criterion

Let me turn now to a second dimension of performance tests--that of test criterion. All tasks have both a product (outcome) and process (steps in task performance). Product measurement is, however, of over riding importance in certifying performance on a task; failure to include

product measurement as the principal criterion may severly limit test validity. Although it may safely be said that every task has a purpose, in practice a great many performance tests employ process measurement only in evaluating a person's readiness to perform outside the classroom.

Before looking more closely at why process measures are widely substituted for measures of task product; we must consider three types of tasks. First, there are tasks in which the product and the process are the same--that is, the product is a process. These tasks are few, and normally serve an aesthetic purpose; examples include springboard diving, dancing, playing a musical composition. Here we see that the product of the task is more or less the correct execution of steps in task performance--that is, the process. Second, there are tasks in which the product necessarily follows from the process. Fixed procedure tasks typically fall in this category. Troubleshooting an electrical circuit, balancing a checkbook, and changing a tire are examples. In such tasks the procedural steps are known and observable, and comprise the necessary and sufficient conditions for task outcome; if the process is correctly executed, task product necessarily follows.

For these first two types of tasks it is not particularly important whether process or product measurement is used. But for a third type, it is very important. This is the type in which the product is not fully predictable from the process--either because we cannot specify all the necessary and sufficient steps in task performance, or because we cannot or do not accurately measure them. In spite of the obvious importance of product measurement for tasks in this latter category, in practice performance tests often do not focus on product. And the reasons generally stem from practical considerations in which the measurement

of task product is viewed as too costly, too dangerous, or too impractical. For example, in a first aid task involving controlling the bleeding from an external wound, the test developer would probably be limited to requiring demonstration of task process; observation of the actual task product--restriction of blood flow--would probably not be possible, for obvious reasons. Other situations are less obvious, however. If any of you are involved in instructor training, you may have observed that a student instructor is evaluated on the basis of such process factors as "had a well organized lesson plan," "used visual-aids effectively," "had good eye contact," "had good voice projection," "covered all points in the lesson plan," and so on. Although the product of instruction is clearly student learning, it is seldom if ever used as the criterion for qualifying an instructor--probably because it would involve a more time consuming method of evaluation.

I'm sure we could all testify to other instances in which product measurement is not used. Some instances are justified by cost or safety considerations; others are not. It seems to me that test developers often fail to see the importance when faced with practical limitations. The overriding question that a test developer should ask himself in this situation is, "If I use only a process measure to test a person's achievement on a task, how accurately can I predict on the basis of this process score whether the person would also be able to effect the product or outcome of the task?" Where the degree of accuracy is substantially less than that to be expected from normal measurement error, the test designer should pause and reconsider how time and resource limitations might be comprised to achieve at least an approximation of product measurement.

Test Conditions

Now, let's look at a third dimension of performance tests: standardization of conditions under which a test is administered. This is an important step in achieving test reliability. Indeed, standardized conditions constitute the very essence of any proficiency measure which professes to be a test. Because this requirement is familiar to test developers, it is seldom violated. Most developers make an effort to maintain test instructions, materials, tools and other environmental factors as nearly constant as possible from one test administration to the next. However, I would like to call to your attention to one particular class of tasks which is particularly troublesome in this regard: tasks involving interpersonal behavior. In such situations, a person or group of persons represents an important part of the environment to be controlled, or standardized from one test administration to the next. Sample situations include counseling, salesmanship, personnel management, or something like hand-to-hand combat. People are part of the task relevant conditions in each of these areas, and obviously people are different to standardize. If you wanted to assess a policeman's ability to properly subdue an unarmed but hostile suspect, what would your performance test be like? How would you insure that test conditions were standardized over all policemen to be tested? The same questions might be asked about assessing a supervisor's ability to persuade a worker to perform some difficult or unpleasant task.

Unfortunately, I know of no easy solution to this problem. Test designers should consider greater use of the well trained, "standardized other." And, here, greater effort should be made to avoid settling too quickly for some probably irrelevant measure of task process.

Test Scoring

The fourth aspect of performance tests I wish to address is test scoring. Scoring protocols primarily affect reliability, but if grossly mishandled in test design--as I will point out in a moment--they may also jeopardize test validity. Scoring procedures involve translating an observed test outcome into an objective pass-fail score. Such procedures should be structured so that only the more reliable perceptual skills are used; that is, the scoring activity should be reduced to one of matching or comparing the test response with some model of correct response. Unfortunately, in many test situations responses seemingly cannot be judged in this "either-or" fashion, but require a "more-or-less" type of judgment. When this occurs the test developer should not (as is sometimes done) compromise by using a test method that yields a more measurable outcome because test validity may suffer. Rather, he should strive to break the task-relevant response down into elements, so that a scorer can more easily make comparative judgments. Typical programs of knowledge testing provide a familiar illustration. The pervasive multiple-choice test yields responses which can be scored with maximum reliability. Scorers obviously have little difficulty in matching a selected response alternative with that which is keyed as correct by the test developer. The scoring of essay tests, on the other hand, has traditionally presented reliability problems. Yet despite the scoring problems inherent in essay testing, a competent test developer would not resort to multiple-choice testing on knowledge tasks demanding recall or generation of material merely to achieve greater scorer reliability. Normally, he would provide a model response in the form of an exhaustive list of the critical elements of an acceptable essay response. The

presence of such elements could then be judged with relative objectivity by a qualified and earnest scorer.

This same thinking applies to the development of scoring protocols for performance tests if these tests are to produce reliable results. The subjectivity with which many task performances are customarily scored could be substantially reduced, it seems to me, through wider use of what may be termed scoring templates. Where the model response on a test of marksmanship is defined as a hole in the bullseye, it is relatively easy for the scorer to judge the acceptability of the response made by the rifleman. The concentric circles normally marked on a target act as a kind of simple template which enhances the ease and objectivity of scorer judgments. Templates could be applied equally well in scoring other tests. For example, tasks mentioned earlier in which the outcome is a process are often difficult to assess reliably. It would appear that performances such as springboard diving or gymnastic exercises could be more objectively scored if the outcomes were filmed and figural templates overlayed on key frames to assess the performer's accuracy at those critical points. Similarly, in evaluating the performance of a music student, recordings of selected renditions could be analyzed at the scorer's leisure--perhaps with the aid of auditory "templates" such as a metronome to measure beat or comparative tones to assess tonal quality. For these particular tasks--or for that matter, any task in which the product is transient--the added cost in recording the product for later scoring would probably be offset by savings in scoring costs; that is, the more objective approach to scoring would very likely preclude the usual requirements for a panel of expert evaluators. But more important, the scorer would not be constrained by real time, and

could function at a place and time and rate of his or her choosing, using prepared templates to increase objectivity.

These four factors--directness of test method, type of performance criterion, standardization of conditions, and objectivity of scoring--must be the focus of further research and creative development work if performance tests are to be used validly and reliably.

Appendix A

Participants in the 1975 National Conference on the Future of Applied Performance Testing

ADMINISTRATORS

Barbara J. Andrew
Director of Research and Development
National Board of Medical Examiners
3930 Chestnut Street
Philadelphia, PA 19104

Henry J. Duel
Management Strategies Associates
2202 Kilt Court
Alexandria, VA 22306

Ralph E. Dunham
U.S. Office of Education
2113 White Oaks Drive
Alexandria, VA 22306

Helen B. Franke
Coordinator of Evaluation Services
Escambia County School Board
5404 Lillian Highway
Pensacola, FL 32506

Hulda Grobman, Discussant
Professor of Health Education
St. Louis University Medical Center
1438 South Grand Blvd.
St. Louis, MO 63104

Mary Hall
Assistant Superintendent
Planning and Evaluation
Oregon State Department of Education
942 Lancaster Drive, N.E.
Salem, OR 97310

Joseph A. Klock
Duval County School Board
Room 11, 1450 Flagler Avenue
Jacksonville, FL 32207

Gerald H. Lunney, Small Group Reporter
Council of Independent Kentucky
Colleges and Universities
Box 668
Danville, KY 40422

Arthur S. McDonald
Director of Research
Nova Scotia Department of Education
Research Section
P.O. Box 578
Halifax, Nova Scotia B3J 2S9

Kenneth R. Olsen
National Learning Resource
Center of Pennsylvania
1509 Woodcrest Circle
Harrisburg, PA 17112

Robert P. O'Reilly
Bureau School and Cultural
Research
Room 481
New York State Education Department
Albany, N.Y. 12224

Ray L. Sweigert, Jr.
Atlanta Assessment Project
1001 Virginia Avenue
Suite 315
Hapeville, GA 30354

CURRICULUM SPECIALISTS

Earl Anderson
Director, Metropolitan Administration Services Center
Multnomah County IED
P.O. Box 16657
Portland, OR 97216

Zita M. Cantwell
Brooklyn College - CUNY
30 West 60th Street
New York, N.Y 10023

Curtis R. Finch
Division of Vocational-Technical Education
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

Craig Gjerde, Small Group Reporter
University of Connecticut Schools of Medicine and Dental Medicine
23 Cassilis Road
West Hartford, CT 06107

William Hulle
Industrial Education
Keene State College
Keene, NH 03431

Kenneth G. Nelson
Teacher Education Research Center
State University Campus
Fredonia, N.Y. 14063

Ruth S. Nickse, Discussant
Coordinator of Assessment
Syracuse University of Research Corporation
Regional Learning Service
405 Oak Street
Syracuse, N.Y. 13203

Gordon E. Samson
Cleveland State University
Cleveland, OH 44106

Charles E. Sherman
Department of Curriculum and Instruction
Illinois State University
Normal, IL 61761

Grady B. Sillings
U.S. Army Signal School
Fort Gordon
325 West Trippe Street
Harlem, GA 30814

John P. Trotta
Curriculum Coordinator
Fern Ridge School District 28J
Elmira, OR 97437

Robert Wenil
Industrial Education
Keene State College
Keene, NH 03431

MEASUREMENT, ASSESSMENT OR EVALUATION SPECIALISTS

James E. Ayrer
Office of Research and Evaluation
School District of Philadelphia
15 Parkside Circle
Winningboro, N.J. 08046

Joseph L. Boyd, Discussant
Educational Testing Service
Princeton, N.J. 08540

Ralph M. Catts
NSW Department of Technical and Further Education
c/o P.O. Box K638, Haymarket
Sydney, NSW Australia 2000

Fred M. Davis
1-A Progress Plaza
Area Learning Resource Center
Harrisburg, PA 17109

Kenneth Epstein
Army Research Institute for the Behavioral and Social Sciences
5610 North 6th Street
Arlington, VA 22205

Gary D. Estes
Phoenix Union High School System
Research and Planning
2526 West Osborn Road
Phoenix, AZ 85017

John M. Finch
Office of Research
South Carolina Department of Education
1416 Senate Street
Columbia, SC 29201

Duane Geiken
DANTES
Ellyson Center
Pensacola, FL 32509

Paul I. Jacobs
National League for Nursing
30 Valley Road
Princeton, N.J. 08540

George O. Klemp, Jr.
McBer and Company
137 Newburg Street
Boston, MA 02116

Joan Knapp
Educational Testing Service
Northgate Apartments 103B
Cranbury, N.J. 08512

Sarah S. Knight, Small Group Reporter
National Assessment of Educational Progress
1953 Ivy
Denver, CO 80220

Marilyn Lichtman
Virginia Polytechnic Institute and State University
Extension and Continuing Education
Reston, VA 22090

Kenneth Majer
Indiana University
104 Maxwell Hall
Bloomington, IN 47401

Jerry Mussio
Assessment Programme
Department of Education
Victoria, British Columbia
CANADA

Navio Occhialini
610 1st Avenue, N.E.
Carmel, IN 46032

Joan Orender
Nebraska State Department of Education
233 South 10th
Lincoln, NB 68508

William C. Osborn, Discussant
Human Resources Research Organization
HumRRO Division #2
Fort Knox, KY 40272

Paul S. Pottinger
Director of Assessment Systems
McBer and Company
137 Newbury Street
Boston, MA 02113

I. Jeffrey Ptaschnik
Stamford Public Schools
Center for Educational Services
Scofieldtown Road
Stamford, CT 06903

Paul Raffeld
Research and Training Center
University of Oregon
565 Harlow Road, #12
Springfield, OR 97477

Jack G. Schmidt
Education Commission of the States
National Assessment of Educational Progress
Suite 700, 1860 Lincoln Street
Denver, CO 80203

Jack Schwille
National Institute of Education
1200 19th Stree, N.W.
Washington, D.C. 20208

Helen Slaughter
Research Department
Tucson Public Schools
District No. 1
1010 East 10th Street
Tucson, AZ 85719

A. J. Stauffer
College of Education
University of Georgia
150 Chinquapin Way
Athens, GA 30601

Richard L. Stiles, Small Group Reporter
Evaluation Section
Office of the Superintendent of Public Instruction (Washington)
812 North Ainsworth
Tacoma, WA 98403

Robert W. Swezey
Applied Science Associates
11316 Links Court
Reston, VA 22090

Margaret A. Wilson
HEW Division of Applied Health Manpower
4615 North Park Avenue
Apartment 1602
Chevy Chase, MD 20015

Appendix B

Handouts Accompanying
The Invited Address

....portant Ideas

for

Executives
Educators
Trainers

on

IMPROVING HUMAN LEARNING

HumRRO

HUMAN RESOURCES RESEARCH ORGANIZATION

To Improve Human Performance

HumRRO, the Human Resources Research Organization, is an independent, nonprofit corporation, whose goal is to improve human performance. In over twenty years of active research, development, and consultation in the area of applied human learning, HumRRO has been a leader in the systems approach to the design, development, and evaluation of effective instruction.

In thumbnail form, this booklet contains some of the main ideas about training and education we have developed. These ideas, we hope, will be useful to those concerned with effective human learning.

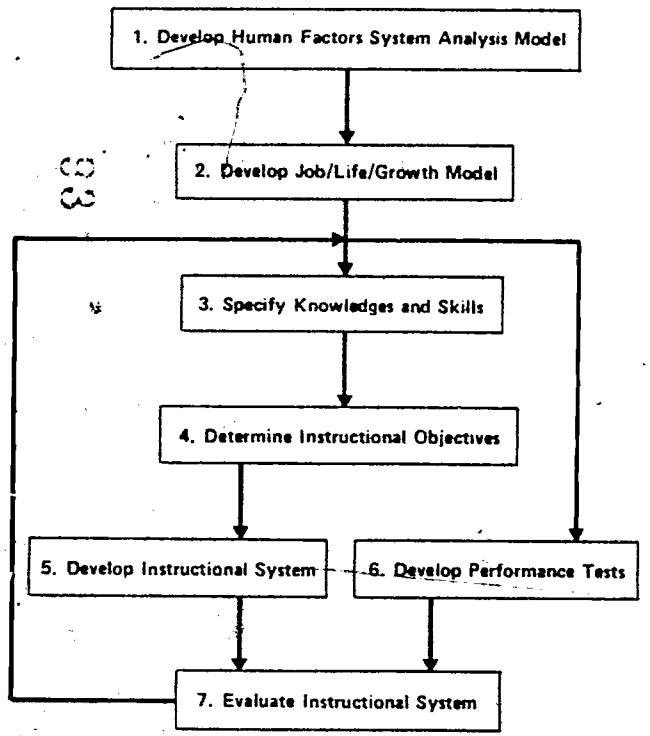
Contents

HumRRO Capability

HumRRO contains broad capabilities to deal with a wide variety of problems concerned with the improvement of human performance and productivity. For more information, contact:

HumRRO
300 North Washington Street
Alexandria, Virginia 22314
(703) 549-3611

HumRRO Model of Instructional System Development



Step 1

Analyze the job or life situation to identify the major needs for human performance.

Step 2

Define the human performance requirements of a job, of a meaningful life situation, or, in the case of education, of growth in more mature behavior.

Step 3

Determine the knowledges and skills which must be acquired by the learner so that he may meet the requirements of the Job/Life/Growth Model.

Step 4

Determine instructional objectives. These are a complete inventory of performances required by the Job/Life/Growth Model.

Step 5

Develop the instructional system itself. This is an integrated set of media, equipment, methods, and personnel efficiently performing the functions required to accomplish one or more objectives.

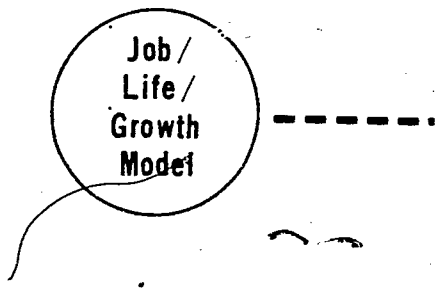
Step 6

Develop tests of performances required by the Job/Life/Growth Model. Note that is a parallel activity to Steps 3, 4, and 5.

Step 7

Evaluate the instructional system, using the tests developed in Step 6. The results are fed back into the earlier stages of the development cycle (line at left).

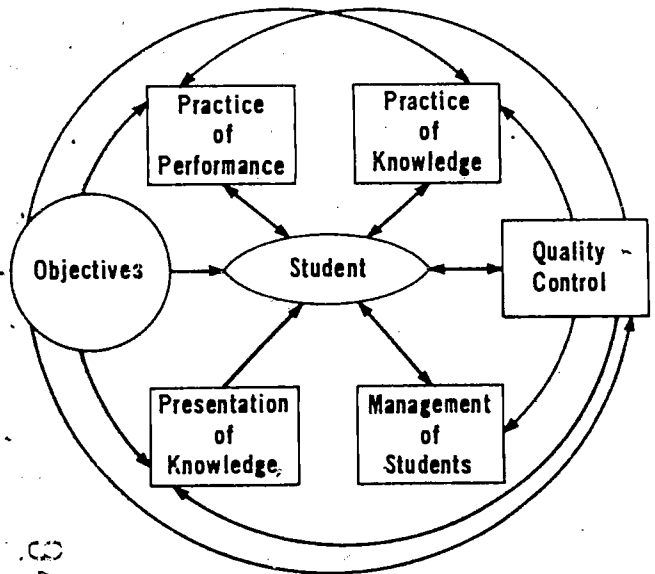
HumRRO Model of an Instructional System



Objectives. Instructional objectives are a complete inventory of performances required by the Job/Life/Growth Model.

Practice of Performance. This function of the system refers to the need to practice whatever performances are required by the objectives. Feedback from the student is important.

Practice of Knowledge. In this function we practice the symbolic processes associated with performance, such as words, pictures, and codes. Feedback from the student is important.



Presentation of Knowledge. A one-way transmission of knowledge from a medium to a student.

Management of students. This function refers to those parts of the system designed to keep the student participating productively in the learning process.

Quality Control. In this function we use tests to determine whether students have met the objectives. The results are used to improve the operation of the system.

A Check List for Evaluating Instructional Systems

This check list contains questions touching on the most important aspects of effective applied learning, both for education and training. It may be used as a guide to evaluation of systems for instruction. It is recognized that measures of student performance are the best ways of evaluating, but this check list can be used when tests are not available. It may also be used to diagnose specific problems with ineffective systems.

Correct Answers are Capitalized

1. Obtaining information concerning the Job/Life/Growth situation toward which learning is directed.

a. Is there a procedure for obtaining information about the Job/Life/Growth situation? YES no

b. Is the procedure applied systematically and consistently? YES no

c. Does the procedure collect performance information for meaningful units of activity? YES no

d. Is performance information actively sought from sources in the work or life situation? YES no

e. Is performance information recorded? YES no

f. Is performance information used systematically and consistently to identify critical instructional needs? YES no

g. Does the procedure provide complete coverage of all likely aspects or occurrences in the work or life performance situation? YES no

h. Does the procedure identify performance actions, conditions, and standards relevant to the work or life situation? YES no

2. Identifying specific instructional objectives.

a. Are decisions about what to teach made on the basis of reliable and valid data? YES no

b. Are detailed analyses made of performance to identify enabling objectives (knowledges and skills required for performance)? YES no

c. Are all enabling objectives required to master a performance identified? YES no

d. Do objectives state precisely the performance actions, conditions, and standards? YES no

e. Do specific objectives use vague terms, such as *know, understand, appreciate, familiarize, general knowledge, working knowledge, qualified?* yes NO

f. Do objectives state what the student will do after instruction? YES no

g. Do objectives state what the course or instructor will do? yes NO

ablishing the sequence of instruction.

a. Is there an effective orientation of the student to the entire job or other life situation in which he is to apply his learning? YES no

b. Are blocks of skills and knowledges taught in isolation from their use in meaningful performance? yes NO

c. Are new skills and knowledges taught only when required in order to master a new meaningful performance? YES no

d. Is the learning of new knowledge followed immediately by practical exercises? YES no

e. Is the relation of each new performance to be learned to the overall purpose of the instruction made clear to the student? YES no

4. Designing situations for the practice of performance.

a. Are practice situations based on an analysis of the behavior to be learned? YES no

b. Does the student practice the entire performance? YES no

c. Has any part of the performance been omitted from practice? yes NO

d. If simulation is used, does it permit practice of the required performance? YES no

e. Have simulations, training devices, or simulators been evaluated to see how well they develop proficient performance? YES no

f. Do instructions for effective use accompany simulations, training devices, or simulators? YES no

g. Have training devices versus actual equipment been subjected to cost-effectiveness analysis? YES no

h. Has the possibility of using obsolete equipment to teach appropriate skills been considered? YES no

i. Do students receive frequent and immediate knowledge of the effectiveness of their practice? YES no

j. Do students receive at least one minute rest between practice trials? YES no

5. Designing situations for the practice of knowledge. YES no

a. Is the knowledge to be practiced clearly related to a meaningful unit of performance? YES no

b. Has information been identified representing the cues to action and the responses the student should make? YES no

c. Has a practice session been planned? YES no

d. Have appropriate practice materials (workbooks, self-instructional programs, flash cards, coaching techniques, etc.) been designed? YES no

e. Do students receive frequent and immediate knowledge of the effectiveness of their practice? YES no

f. Do students maintain a record of their progress during practice? YES no

6. Preparing presentations to the student.

a. Has the content of the presentation been tested on students to determine, by means of achievement tests, whether it communicates to the students? YES no

b. Is the content of the presentation meaningful to the student? YES no

c. Are there lengthy periods of presentation uninterrupted by practice? yes NO

d. Are films and television integrated with practice? YES no

e. Are lectures, demonstrations, films, books, or tape recordings selected on a cost-effectiveness basis? YES no

f. Have texts and other reading matter been examined to be sure they are within the reading capability of the student? YES no

7. Maintaining student learning activity.

a. Has the degree of spread in aptitude scores of the students been determined? YES no

b. Have adjustments been made to the instructional program to accommodate students of different aptitudes? YES no

c. Have the student's interests, educational background, and attitudes toward formal schooling been determined? YES no

d. Is this information used to make instructional presentations more meaningful to the student? YES no

e. Do all students receive rewards, praise, or opportunity to engage in preferred activities, when they achieve course objectives? YES no

f. Are successful students treated in ways that represent punishment to them? yes NO

g. Are failing or borderline students treated in ways that represent rewards to them? yes NO

8. Control of the quality of instruction.

a. Are the tests direct translation of the objectives? YES no

b. Is emphasis given to performance tests? YES no

c. Are grades expressed in percentage passing? YES no

d. Are grades based on the normal, bell-shaped curve? yes NO

e. Are grades based on percentile ranks? yes NO

f. Are test items changed to make them easier, or harder, or to conform to an "ideal" distribution of grades? yes NO

g. Are results of student testing provided to the instructional departments? YES no

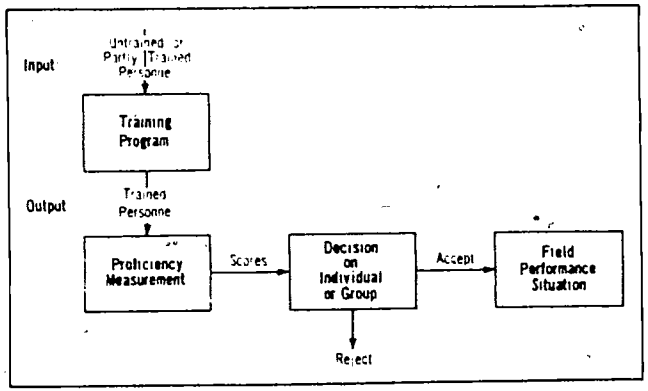
h. Do the instructional departments make changes in procedures, media, instructional methods, and content suggested by the results of student training? YES no

i. Is the quality control program under the control of a unit separate from the instructional departments? YES no

Evaluation—Accountability—Quality Control

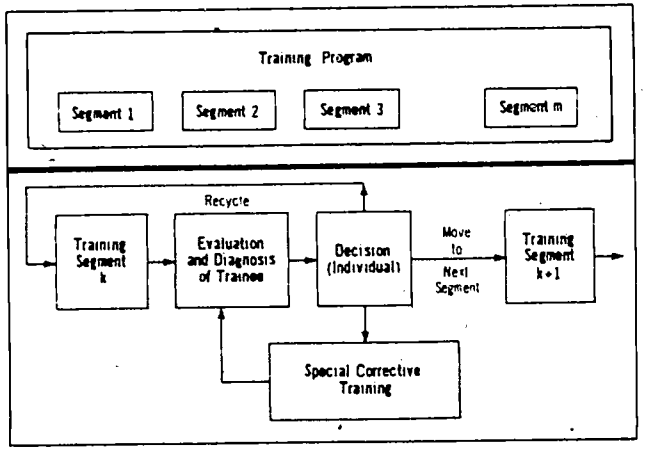
There are four specific objectives of systems of evaluation, accountability, and quality control:

1. Quality Assurance



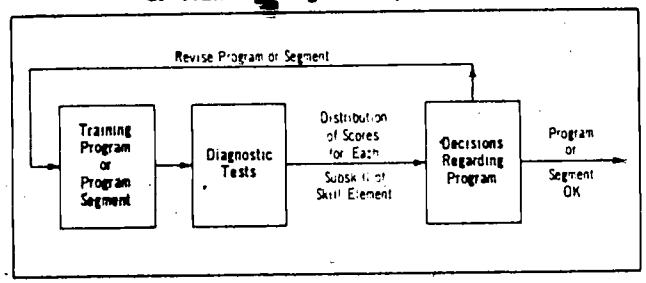
- Students or personnel enter instructional program
- Trained students or personnel leave program
- Proficiency measures applied
- Decision to reject or accept into job/life situation based on test scores

2. Control of Student Progress



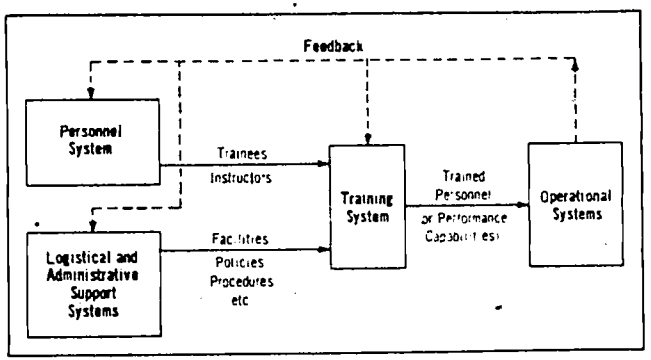
- Instructional program is composed of segments (see upper half of figure)
- Now refer to lower half of figure
- Decision to recycle, give corrective instruction, or move to next segment based on tests

3. Training Program Improvement



- Students leaving program or segment are given tests
- If scores are satisfactory, the program is left as it is
- If scores are not satisfactory, the program is strengthened

4. Training System Diagnosis



- Instructional system is related to other systems
- Perhaps a change in other systems is the way to improve the instructional system output
- Suitable measurements can improve system relationships

Selected References

If you wish to learn more of the details of the topics sketched here, the following references are suggested:

Ammerman, Harry L. and Melching, William H., *The Derivation, Analysis, and Classification of Instructional Objectives,* HumRRO Technical Report 66-4, May 1966.

Cogan, Eugene A., Hoehn, Arthur J., and Smith, Robert G., Jr., *A Framework for Viewing Quality Control in Training,* HumRRO Professional Paper 28-70, November 1970.

Haggard, Donald F., Willard, Norman Jr., Baker, Robert A., Osborn, William C. and Schwartz, Shepard, *An Experimental Program of Instruction on the Management of Training,* HumRRO Technical Report 70-9, June 1970.

McClelland, W.A., *R&D: What Industry Can Learn From Research in Army Electronics and Electrical Maintenance Training,* HumRRO Professional Paper 2-69, January 1969.

Olmstead, Joseph A., *Theory and State of the Art of Small-Group Methods of Instruction,* HumRRO Technical Report 70-3, March 1970.

Smith, Robert G., Jr., *A Manpower Delivery System: Implications for Curriculum Development,* HumRRO Professional Paper 19-70, June 1970.

Smith, Robert G., Jr., *The Engineering of Educational and Training Systems,* Lexington, Mass., D.C. Heath, 1971.

Taylor, John E., Montague, Ernest K., and Michaels, Eugene R. *An Occupational Clustering System and Curriculum Implications for the Comprehensive Career Education Model,* HumRRO Technical Report 72-1, January 1972.

Taylor, John E., Smith, Robert G., Jr., *The General Concept of Managing for Accountability,* HumRRO Professional Paper 4-72, February 1972.

Yagi, Kan, Bialek, Hilton M., Taylor, John E., and Garman, Marcia, *The Design and Evaluation of Vocational Technical Education Curricula Through Functional Job Analysis,* HumRRO Technical Report 71-15, June 1971.

Professional
Paper
16-72
HumRRO-PP-16-72

HumRRO

Frameworks for Measurement and Quality Control

Eugene A. Cogan and J. Daniel Lyons

Presentations at
New York University
First National Annual
Training in Business and Industry Conference
New York City
March 1972

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street • Alexandria, Virginia 22314

July 1972

The Human Resources Research Organization (HumRRO) is a nonprofit corporation established in 1969 to conduct research in the field of training and education. It is a continuation of The George Washington University Human Resources Research Office. HumRRO's general purpose is to improve human performance, particularly in organizational settings, through behavioral and social science research, development and consultation.

Published
July 1972
by
HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street
Alexandria, Virginia 22314

BIBLIOGRAPHIC DATA SHEET	1. Report No. HumRRO-PP-16-72	2.	3. Recipient's Accession No.
4. Title and Subtitle FRAMEWORKS FOR MEASUREMENT AND QUALITY CONTROL			5. Report Date July 1972 6.
7. Author(s) Eugene A. Cogan and J. Daniel Lyons			8. Performing Organization Rept. No. PP-16-72
9. Performing Organization Name and Address Human Resources Research Organization (HumRRO) 300 North Washington Street Alexandria, Virginia 22314			10. Project/Task/Work Unit No. 11. Contract/Grant No.
12. Sponsoring Organization Name and Address			13. Type of Report & Period Covered Professional Paper 14.

15. Supplementary Notes
Two presentations at New York University First National Annual Training in Business and Industry Conference, New York City, March 1972.

16. Abstracts The author of the first paper states that, in making a job performance evaluation, anything that can be specifically defined can be measured. However, to develop a testing program that is both useful and cost-effective, it must be known who will make what decisions, using the obtained measurements. Analysis and interpretation of the particular pupose and setting are needed. Feedback data show how improved decisions can produce dollar gains far beyond the cost of developing and employing measurement. In the second paper, the essential elements of a quality control system are illustrated, including (a) training objectives or performance requirements, (b) proficiency and diagnostic measures, (c) data reduction and analysis, (d) procedures for decision and corrective action, (e) communication procedures, and (f) managerial support. It is shown that training goals must be defined in terms of measurable on-the-job performance.

17. Key Words and Document Analysis. 17a. Descriptors
*Performance evaluation
*Performance standards
*Performance tests
*Measurement
*Test construction
*Validity
*Quality Control
*Job analysis
*Personnel selection
*Work measurement
*Training systems

17b. Identifiers/Open-Ended Terms
Quality control systems
Instructional systems
Accountability

17c. COSATI Field/Group 05/09

18. Availability Statement	19. Security Class (This Report) UNCLASSIFIED	21. No. of Pages 15
Distribution of this document is unlimited.	20. Security Class (This Page) UNCLASSIFIED	22. Price

FORM NTIS-35 (10-70) USCOMM-DC 40329-P71

Prefatory Note

These papers were presented at the First National Annual Training in Business and Industry Conference of New York University, held in New York City in March 1972. The first paper, " 'If It Exists, It Can Be Measured'—But How?" was prepared by Dr. Eugene A. Cogan, who is Director for Research Design and Reporting in the Executive Office of the Human Resources Research Organization (HumRRO) in Alexandria, Virginia. The second paper, "Measuring Effectiveness: Quality Control of Training," was prepared by Dr. J. Daniel Lyons, who is Director of HumRRO Division No. 1 (System Operations), also located in Alexandria, Virginia.

"IF IT EXISTS, IT CAN BE MEASURED"—BUT HOW?

Eugene A. Cogan

Psychologists—including those especially interested in measurement—have been, and continue to be, plagued by elusive and fragile concepts. Many concepts have their origin in the individual and cultural experiences all people share. For example, we all have the feeling that we know some people who seem "smarter" than others over and above differences in their schooling or other educational experience; this feeling has led to the concept of "intelligence" and to attempts to define, understand, and measure intelligence. Our shared experiences have led us to feel that some people are better employees than others; this feeling has led to attempts to define, understand, and measure "goodness as an employee." Attempts to cope with "goodness as an employee" have been equally as frustrating to employers and to psychologists as have been attempts to make sense of "what is intelligence all about."

The main stem of the title of my paper—"If it exists, it can be measured"—is a free translation of a classic statement by Edward Thorndike who was trying to counter the pursuit of poorly defined pseudo-concepts that bordered on being personal illusions. For us, Thorndike's message is: "Until you can define what you are interested in well enough so that you can figure out how it can be measured, it can mean anything and, therefore, it means nothing."

The challenge of Thorndike's proposition to theoretical psychology has no easy answer because theoretical psychology *is* concerned with generally important abstractions regarding human behavior. There is an understandable reluctance to fix on formal definitions for concepts because useful definitions must be restrictive and omit things: theoretical psychologists are reluctant to risk throwing out a baby with the bath water.

However, for practical, applied measurement the implications of Thorndike's doctrine are very useful. In a practical setting, Thorndike's edict translates to: "Of course you can measure it, after you have defined what it is." The main purpose of my presentation will be to deal with how to go about defining "it" so that you can proceed to measurement, and then how to evaluate the measurement.

In any practical setting, there are many situation-specific features and these provide a key to measurement. The trick to translating an impression into a measurable something consists of using the situation to define what measurement is needed.

Purpose of Measurement

Foremost for defining measurement is "why." In selecting or devising a measurement, it is essential to decide or determine the purpose of the measurement. In industry, the purpose translates to decisions that management or personnel people must make. Who will decide what with the aid of measurement information?

It is not enough to stop analysis of purpose at the broad levels of selection, assignment, promotion, training evaluation, or personnel evaluation. Each of these includes so many variants that depend on *particular* purposes that the category is the beginning, not the end, of analysis. If concern is with selection, the proper measurement depends on whether selection is for training or for direct job assignment, whether concern is solely for competence in an entry job or also with potential for advancement,

whether the work setting is closely supervised or relies on self-supervision, whether the work setting requires team work or individual work, and so on.

Even what seems to be a specialized and highly specific purpose like quality control of training, as is shown in Dr. Lyons' paper, involves at least four distinctive purposes and each of these has its own distinct definition and measurement.

What is Measured

In a particular setting, with purpose established in terms of the particular decisions that are to be made, the second element in defining the measurement concerns what is to be measured. Much of the definition of "what" will already have been established in careful definition of purpose. That is, if the purpose concerns selection for a training program preceding assignment to a job, the "what" should not contain very many, if any, direct indications of job knowledges and skills, but rather should deal with ability to learn these knowledges and skills. On the other hand, if selection is for direct assignment to job duties, it is whether these have been previously learned that is pertinent.

The matter of what is to be measured has been, by far, the subject of most concern and debate in industry and among measurement specialists. Primarily, this is because dollars and time for measurement, cost elements that are very sensitive in industry, are heavily dependent on what is measured. For example, considering job performance evaluation, the best theoretical measure is unobtrusive, scientific observation and careful measurement of behavior, over a long period of time, in the actual job setting. While such measurement is technologically possible, it would be so prohibitively expensive that less costly alternatives are always being sought and, typically, used. However, these less costly methods do not measure the same thing.

Usually considered closest to scientific observation in the natural setting is a job sample test. Even assuming that sampling of the job performances is well done, job sample simulation is not the same as job observation because important contextual and personal elements cannot be simulated. That is, a test environment creates test performance for the individual. He may try much harder than he does in the natural setting, or he may be immobilized by test anxiety.

Less costly—and hence more common than job sample simulation tests—are analytic tests of job performance elements. Such tests measure component skills and knowledges underlying job performance. We are all familiar with such analytic tests as they apply to selecting a secretary. For a candidate secretary, one might use a typing test, a dictation test, and a spelling test. While such tests can provide assurance that necessary individual job skills are within the candidate's repertoire, they do *not* assure the person can fit the skills and knowledges together effectively in a job setting, or that the person can or will do the many other tasks required on the job.

For still less cost than analytic tests, there are indirect tests of capabilities, usually paper-and-pencil tests dealing with incidental information about the job.

The simplest of the indirect tests are specialized vocabulary tests. For example, a good secretary is likely to know what "platen" means, and what a number four pencil is, and what the term "stay-back file" means. Since none of these three items of information is intrinsically of consequence in doing a good job as a secretary, they constitute *indirect* measures.

Use of indirect measures must be approached with great caution and checked empirically against more direct measures. This is because possessing such information may not come from job competence—witness the fact that I know the meaning of the three terms, but I have no secretarial competence whatsoever.

Most common of all as a measurement of job performance in industry is the rating scale. The reasons are that, first, it is the least expensive measure and, second, it seems to make sense to go to the day-to-day observer of job performance who has "seen job

performance with his own eyes over a long period of time." Despite the sensibility and low cost of rating scales, they don't do what most people think they do. Rating scales—regardless of what the rater is asked to check—provide a measure of an overall "Joe is OK by me," rather than how well Joe can perform elements in his job. I do not at all intend to deprecate the value of personnel decisions based on "Joe is OK by me"; I wish, however, to emphasize that what is being measured in that fashion differs from what is measured by a performance test even if the terms used are similar.

There are differences in what is being measured for all the categories named: natural observation, job sample tests, analytic tests, indirect tests, and rating scales. Treating them as alternate techniques to measure the same thing can be severely misleading. It is traditional to consider these measurements as alternatives, differing in technique but not in what is being measured. This inaccurate assumption of equivalence is possible only because not enough—and not precise enough—analyses have been performed to define purposes of measurement and what is to be measured.

Effectiveness of Measurement

I will now turn to effectiveness or—in psychometric terms—validity, as it applies to the consideration of measurement.

I began this paper with the proposition that one first must define carefully and analytically the precise purpose of measurement, taking into account the organization setting; then I pointed out that purpose translates to who will make what decision using the measurements. Second, I proposed that purpose and decision should be the key ingredients in determining what will be measured, but I only touched on how one goes about translating purpose into what is measured. I skirted the transition because only gross and tentative rules or guidelines are available. Basically, the measurement specialist must—as a first cut—use his best judgment. Since his best judgment may be wrong or may be severely distorted by cost or other practical considerations, it is essential that the development of a testing program be viewed as a cyclic feedback process, or a cut-and-fit process, with a continual flow of information on whether decisions using test data are good ones. Information on the flaws in such decisions provides the means for changing the measurement and—over time—shaping measurement to maximum support of the decisions that need to be made.

The term "validity" in psychology has many meanings—and the meaning varies depending on the person and on the context in which the term is used. For this reason, I shall avoid these ambiguities and discuss more broadly what one should consider in dealing with the effectiveness of measurement.

The first question to consider is the accuracy of the measurement. What are the tolerances of the emerging numbers?

It is tempting to propose "the more accurate the better." But, that proposal is untenable because cost of measurement increases as requirements for precision increase, in the same way as measurement to one-ten thousandth of an inch is more expensive than measurement to the nearest foot. Just as we decide on tolerances for a length measurement by considering our purpose—whether it is watch-making or road-building—the precision needed in psychological measurement depends on the purpose of measurement, that is, the nature of the decision that is to be made.

The second question regarding effectiveness of measurement concerns stability. If one retested at some later time, how similar would the measurement numbers be to a first set of numbers? Psychologists normally call this characteristic "reliability" but, as with the term "validity," "reliability" has multiple meanings and use of the term is more likely to confuse than to clarify.

How much stability is needed? The hoary tradition of psychological measurement includes the rule that a "correlation of .8 or more is needed for individual decisions; a

correlation as low as .3 can be used for group decisions." This serves as a *general* rule of thumb and, therefore, cannot fit anything. Much better than the all-fitting and hence never-fitting rule is the analysis of purpose and what is to be measured. From analysis of the purpose, one can define the kind of stability of measurement that is needed. From analysis and interpretation of what is being measured, one can distinguish between stability of measurement as it pertains to mechanics of measurement and as it pertains to the nature of what is being measured. In some instances, stability over time would be nonsense, for example. Suppose we administer a typing proficiency test to a group about to begin training in typing. Wouldn't it be foolish to expect test scores secured after training to be about the same as the first set?

The third question under the heading of effectiveness is the pay-off. How much better, in practice, are the individual decisions reached using the measurement than those reached without such information? This question can readily be cast into terms very familiar in industry: How much would it cost to save how many dollars? What is the net gain? However, in order to do such an analysis, it is absolutely necessary—to revert to my main thesis—that the purpose of measurement be analyzed and defined very explicitly, down to exactly what decisions will who make using the measurement data. With decisions defined, it is possible—and, perhaps, even routine—to perform a cost-effectiveness analysis of psychological measurement.

Measurement in industry has enjoyed only mixed success at best, and the question "Is testing worth it?" addressed to management most often results in the answer "I don't know." I think there are two related reasons for this unclear state of affairs.

First, there are many industrial managers who enter internal, deliberative policy councils with a personal conviction that what is really important cannot be measured by tests and that tests and psychologists are not to be taken seriously. In that same council, frequently, will be a testing enthusiast and, after a period of wrangling, the traditional compromise will occur: "Let's try it out on a small scale." Unfortunately, the small-scale approach frequently leads to skipping the crucial steps of analyses to establish purposes, to the level of who will make what decisions with the information. Therefore, any hope of getting a good fix on exactly what is to be measured is sacrificed. Usually, a conveniently available test with a name that seems about right and that may have been recommended as a good test is chosen for trial purposes—whether or not it fits the situation and purpose.

Second, exacerbating the instant magic of choosing a convenient test is the fact that, rather than programing a systematic cut-and-fit program for choosing and/or developing measures, a one-shot tryout is undertaken. If the test passes, it's in; if not, testing is out for the company.

Good testing is more expensive than poor testing or no testing. Analysis to determine whether good testing is worth the trouble is not very difficult, once analysis and definition have proceeded to the level of who will make what decision with the information. The costs of poorer decisions in excessive training costs, reduced productivity, or costs of firing someone and hiring a replacement can be estimated, at least roughly. In addition, costs of developing and using a measurement system can also be estimated, at least roughly. From such data, one can calculate a break-even point in terms of the amount of improvement in decisions that is needed to recover costs of measurement. Usually, since training, selection, hiring, firing, and other consequences of decisions are so very expensive, it will be found that even miniscule improvement in the quality of decisions will more than pay for a good measurement program.

Summary

In closing, I should like to repeat my main points:

First, philosophical disputes about whether a person's characteristics can be

measured are pointless. Anything that can be specifically defined can be measured. Such definitions should be in terms of behaviors that can be observed.

Second, to develop a testing program that is useful and cost-effective, the planned use of the test information must be carefully defined: that is, who will make what decisions using the measurements to be obtained.

Third, analysis and interpretation of the particular purpose and the particular industrial setting are essential to decide, hypothesize, estimate, or guess what should be measured. What are usually considered to be different measuring techniques for the same thing are, in fact, measures of different things.

Fourth, the effectiveness of measurement should be evaluated in terms of precision stability, and amount of improvement in organizational activities, all of these considered in terms of the decisions for which measurement provides support. For maximum return on the testing dollar, it is essential to proceed cyclically, continually improving the measurement program in the light of feedback on how decisions are improved—or not improved—by measurement data.

Fifth, analyses of saving that can be accomplished by improved decisions are usually startling, producing dollar gains far beyond the cost of developing and employing measurement.

My main thesis has been that measurement must be considered in the particular framework in which it is to be used—and here I am talking about measurement in general! I, therefore, call your attention to Dr. Lyons' presentation on quality control, an excellent illustration of the concept of defining who will make what decision using what measurement information.

MEASURING EFFECTIVENESS: QUALITY CONTROL OF TRAINING

J. Daniel Lyons

As the philosopher Seneca observed, "When a man does not know what harbor he is making for, no wind is the right wind." And when training goals have not been precisely defined in terms of measurable on-the-job performance, no training technique is the right training technique. The most pervasive weakness of training programs is lack of precision in locating the harbor of improved job performance. As a result, they are buffeted constantly by the winds of promise and innovation in training—but no wind is the right wind.

Development of new training programs and the introduction of changes in existing programs are fruitless exercises unless and until the means for assessing progress toward precisely defined goals have been developed. Behavioral psychologists have been portrayed by some critics as "drab purveyors of the obvious." In this paper, I may well be adding credence to that observation. It is obvious, is it not, that one does not introduce change unless there exist mechanisms for assessing the effect of the change? I am in the role of a drab purveyor of that obvious and fundamental principle. Because in government, industry, the public schools, and wherever training and educational programs exist, that obvious principle is being continually violated—at a fantastic cost in wasted dollars and human potential.

The process of developing the raw material of human potential deserves a system of quality control at least as carefully developed as that applied to the manufacturing process. By a quality control system I mean essentially an information system and a system of concepts, models, and procedures designed to accomplish four main objectives:

(1) Quality assurance
(2) Control of student progress
(3) Training program improvement
(4) Training system diagnosis and change

The quality assurance function is illustrated in Figure 1.

Does the product meet the specifications? This question cannot legitimately be posed unless and until the specifications have been delineated in terms of operational requirements and these requirements have been reflected in end-of-course proficiency measures. The intent is to rid the training system of criteria based on *amount* of training in favor of demonstrated proficiency in the required job elements. Systematic application of precise job performance criteria through a quality control system results not only in an improved product, but also in the discarding of irrelevant material. Thus, the cost of installing an effective quality control program is amortized through savings in the training program, particularly in personnel time of instructors and students.

The second objective of a quality control system is to provide a means of selecting and organizing the learning experiences of the students to facilitate achievement of the objectives.

The training program depicted in Figure 2 is composed of a series of segments or modules (upper half, Figure 2). Conceptually, these may be as long as a major phase of the course, or as short as a single brief lesson. Each such segment or module is designed to help the student meet specified learning objectives.

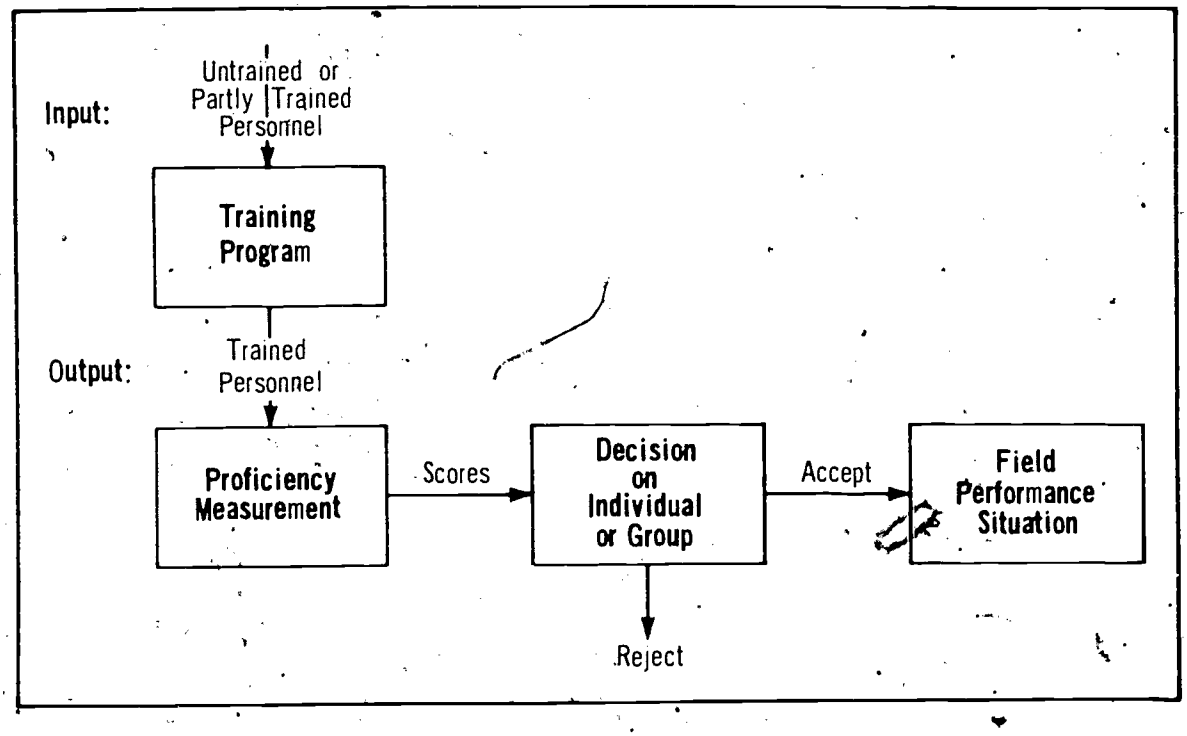


Figure 1

The decision options (lower half, Figure 2), include those of sending the student forward to the next segment of the course, recycling, or giving special corrective training. Generating information to aid in choosing among the options is a function of a quality control system. It should be noted that the option of special corrective training is contingent upon the precision of the diagnostic instrument; that is, the evaluation procedure must be capable of identifying specific weaknesses toward which the corrective training can be directed. The goal is a system by which the trainee is continuously evaluated, selectively corrected, and advanced as performance standards are met, and *only* as they are met.

The first two objectives, quality assurance and control of student progress, are concerned with assessment of student performance. The third objective, shown in Figure 3, is that of program improvement; the emphasis is on program assessment rather than assessment of the individual trainee. Unfortunately, too often changes in training programs tend to be based on administrative edict. We are all familiar with those frustrating situations in which changes in management bring about changes to conform to the biases of the new manager; for example, the shifting emphases on theory and practice in the training of repairmen depending upon the views of upper management rather than job requirements and performance. A systematic quality control process that can identify weaknesses and strengths in the program by assessing and diagnosing the performance of the trainee provides a bulwark against the shifting winds of administrative edict. Further, the control process is necessary in order to assess the effects of changes made to strengthen the program. The most important motivator that can be supplied to any trainer is precise and accurate feedback on the results of his efforts. If this is supplied, training *will* improve, if only by trial and error.

Control of Student Progress

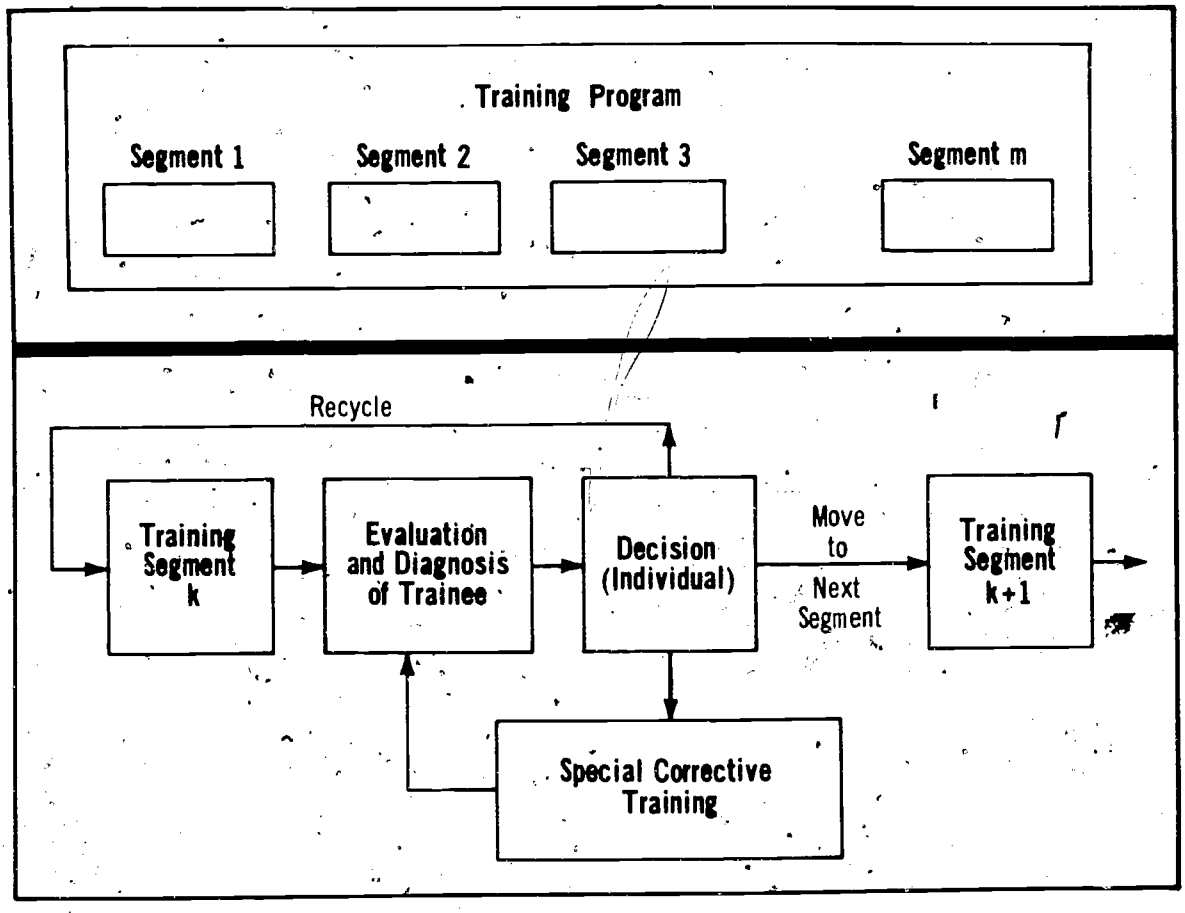


Figure 2

Training Program Improvement

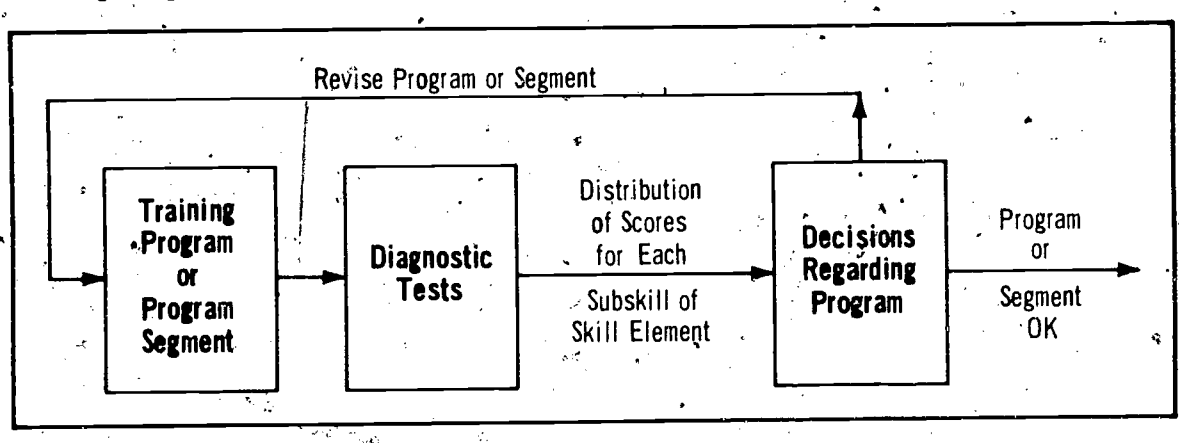


Figure 3

From a Training Director's point of view, Figure 4 may be viewed in the following manner. From the operational elements of the organization, the training system receives performance requirements that are ridiculously inflated or impossibly vague, which must be met with trainees and instructors of minimal aptitude and experience supplied by the Personnel Department, while operating under policies and procedures that are unrealistic, or inflexible, or antiquated, or obscure, or all of these, while utilizing outdated equipment and facilities, and operating on a miniscule budget.

Training System Diagnosis

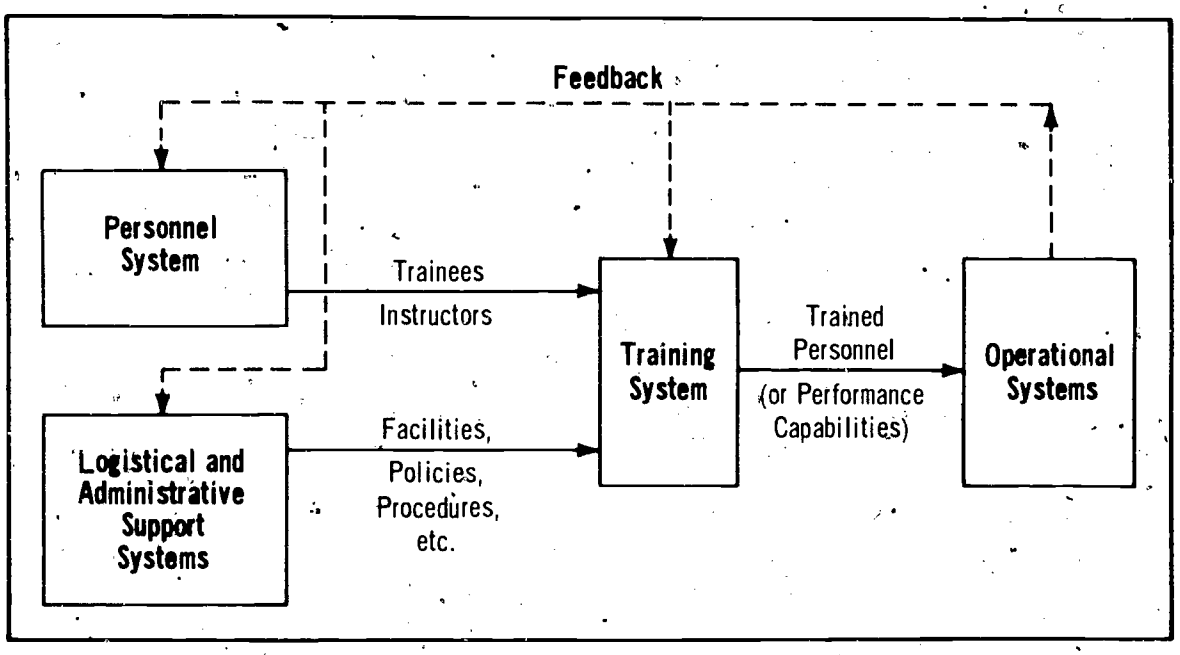


Figure 4

While that may be the world as seen by those of us concerned with training, it is safe to assume that each of the other elements of this system *and* management see somewhat different worlds. An adequate quality control system can alleviate the resulting stresses and strains by providing the information that helps to identify and define the problems and to evaluate the effects of attempted solutions.

The training system is all too often the scapegoat for problems resulting from poorly defined operational requirements, inappropriate utilization of training system products, inadequate personnel selection procedures, and ill-conceived personnel policies. A well-designed quality control system can serve to put our training house in order and provide the basic information for productive interaction with other systems in the organization. In short, it may get the monkey off our backs *or* fasten it there more firmly, if appropriate.

The essential elements of a quality control system are:

(1) Training objectives (performance requirements)
(2) Proficiency and diagnostic measures
(3) Data reduction and analysis
(4) Procedures for decision and corrective action
(5) Communication procedures
(6) Managerial support

For quality control, crucial information derives directly from training objectives. They form the keystone for a useful and effective quality control system by providing not only the specifications for instruction, but also the basis for evaluating instruction. Thus, we must begin with a complete set of good training objectives for a training program, and these objectives represent the mission of the training system.

Management plays the beginning role with regard to training objectives by defining exactly what is to be accomplished by the training system. The raw material for such defining comes from many sources—policies, plans, specifications for new equipment, information concerning on-the-job performance of earlier graduates, information about on-the-job requirements, and so forth.

The management element assembles all such information and decides on *terminal* training objectives. In order for the *terminal* objectives to be most useful, they should be in the form of detailed specifications.

With terminal objectives defined, the training operations element is responsible for developing detailed training objectives and for providing graduates who can perform as defined by management. The set of terminal objectives forms a complete inventory for evaluation. The training objectives also include information about the conditions under which tasks are expected to be performed and thereby define test conditions. Further, the training objectives also include the standards or tolerances for the tasks in terms of accuracy and speed requirements; these are also tolerances for use in scoring an individual's performance on a task.

In order to assess the effectiveness of how the training system is performing, another kind of information is needed about each task—the minimum acceptable percentage of students capable of performing within tolerances. Cost and time aside, it would be desirable for every student to be able to perform every task within the defined tolerances. However, achieving such a goal would be likely to make the cost and time for training intolerably large. Something short of 100% of the students capable of 100% of the tasks must be defined as an acceptable standard of effectiveness of the training system.

The standard must, however, take account of the varying criticality of the tasks. Ninety percent of electricians being 90% correct in the procedures for grounding an electrical circuit during repairs is *not* an acceptable standard. Fifty percent knowing the correct nomenclature of 50% of the contents of their tool kits may be acceptable on a particular job. The criticality measure for any task is basically an assessment of the effect on the operational system of the incorrect performance on that task. In assisting in the development of a training program for stock clerks, we found that the system could absorb, with minor turbulence, an error in the nomenclature of an ordered item but that the stock *number* was highly critical—a misplaced digit could produce an avalanche of toilet paper instead of a fork-lift truck. Similarly, the delivery address was of medium criticality, producing serious delay in delivery—but a misreading of the unit of issue—and we have an avalanche of toilet paper.

The second element, tests and measures, does not make a quality control system—yet they are clearly an essential element of any such system in order to provide the data base on which the system rests. In quality control we are particularly concerned with the diagnostic capability of our testing procedures. We must be able to pinpoint the strengths and weaknesses of the training for each detailed objective as a basis for decision and action to improve or modify the training. In the light of Dr. Cogan's comprehensive discussion of tests and measures, further discussion of this topic seems unnecessary.

It should be re-emphasized, however, that quality control requires absolute rather than relative criteria. Scores and grades must reflect how many of course objectives have been mastered rather than how a student compares with other students. Further, we must ensure that we are not wasting our training time and the potential of our trainees by

failing them for the wrong reason. The key is job-relevance of both training and testing. If the job requirement is to replace the bad part in a TV set on the basis of observation of symptoms, the ability to quote and manipulate Ohm's Law is not job-relevant. Our carefully controlled studies document the fact that many potentially excellent electronics repairmen in a number of training programs have been discarded because of irrelevant weaknesses in physics and mathematics.

The test scores in and of themselves carry little meaning. As a third element, test data must be analyzed and interpreted before they can yield meaningful inputs to decision processes. The data reduction generally involves three kinds of considerations—central tendency, variability, and stability. The central tendency is calculated to show the overall performance of the group—average, mean, or perhaps, more useful, the percentage of a class able to perform each specific task at or above the minimum standards. The variability or spread is generally characterized by calculating the standard deviation, while stability is identified by the standard error in order to distinguish the accidental or incidental deviations from those that have a "real" basis.

In the analysis of the data that have been reduced to measures of central tendency, variability, and stability, three basic questions arise regarding performance on each task. First, how does the central tendency compare with the standard? Has the class performed above, below, or at the standard? Second, does the class performance fall within tolerances established for the standard? Third, how critical or important is the task to operational performance? As indicated earlier, the criticality of the task has direct implications for the urgency of corrective action. The criticality dimension is built into the analysis by differential standards and tolerances for specific tasks.

The collection, reduction, and analysis of the test and performance data are necessarily designed to support a program of corrective actions, the fourth essential element of the quality control system. It is, unfortunately, almost commonplace to find massive collections of training data, created at considerable effort and expense, lying idle. Too often such data are assembled without a specific plan for utilization or in the absence of specific procedures for implementing the existing plan. Prior to the collection and analysis of the data, there must be procedures for corrective action—that is, specification of the process by which decisions are made and means of assigning responsibility for implementing the actions selected. These procedures should be designed to identify problems and to assign priority to their solution. The highest priority for action is for those cases where the data analysis shows that performance is seriously out of the tolerance range.

In order to maintain confidence and support of management and of the operating elements, it is important that such problems be identified by the training element and corrective action initiated immediately. The system should act rather than react to external complaints. A complete action program should include procedures for:

(1) Identifying points and places where something seems to be seriously out of tolerance and immediate action is indicated.

(2) Identifying points and places that are "suspicious," and that warrant investigation as time and resources permit.

(3) Establishing a normal routine work load for continuing study of the training program when everything is going well.

Obviously a quality control system must include carefully designed communication procedures. The information generated by the system must be differently packaged for transmission to the responsible individuals on an appropriate schedule so that the necessary decisions can be made on a timely basis. Equally important are provisions for flow of relevant information into the system—changes in operating procedures, new equipment, modifications in personnel selection procedures, policy decisions affecting training, and so on.

Proper communication is vital to maintaining managerial support, which is both a cause and an effect of a dynamic quality control system. The quality control system cannot operate effectively without strong support from all managerial levels, nor will this support continue unless the system operates effectively. Support from management is especially needed, because the data produced by the quality control element may be unpleasant. However, if the information is directed toward corrective action, quality control can be viewed as the shared mission of management and the training element: producing the tangible asset of a well-trained addition to the company work force.

Professional
Paper
3-73
HumRRO-PP-3-73

HumRRO

Developing Performance Tests for Training Evaluation

William C. Osborn

Presentation at
U.S. Continental Army Command
Training Workshop
Fort Gordon, Georgia October 1971

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street • Alexandria, Virginia 22314

February 1973

Prepared for

Office of the Chief of Research and Development
Department of the Army
Washington, D.C. 20310

The Human Resources Research Organization (HumRRO) is a nonprofit corporation established in 1969 to conduct research in the field of training and education. It is a continuation of The George Washington University Human Resources Research Office. HumRRO's general purpose is to improve human performance, particularly in organizational settings, through behavioral and social science research, development, and consultation. HumRRO's mission in work performed under contract with the Department of the Army is to conduct research in the fields of training, motivation and leadership.

Published
February 1973
by
HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street
Alexandria, Virginia 22314

Prefatory Note

This paper was presented at the U.S. Continental Army Command Training Workshop at Fort Gordon, Georgia in October 1971. The research on which this paper is based was performed under Work Unit TRAINMAN, Development of an Instructional Program in Training Technology and Training Management, at the Human Resources Research Organization, Division No. 2, Fort Knox, Kentucky.

DEVELOPING PERFORMANCE TESTS FOR TRAINING EVALUATION

William C. Osborn

A performance test is a template—a template modeled from a job task and used to gauge the similarity of a trained behavior to the demands of that job task. This view of performance tests implies a straightforward approach to their development. One simply re-creates the circumstances of the job task, asks the trainee to perform the task, and then records that he did or did not do it. Unfortunately, from our own experience we know that it is not this simple. Many practical problems intervene to complicate the process. We often find that a job has so many tasks that days would be needed to test them all. Occasionally, the equipment, terrain, and other support requirements prevent a realistic test for even a single task. At other times, we run into standards of task performance that are difficult to translate into a pass-fail criterion for scoring. We also have found that trainers need more than pass-fail results; they need diagnostic information to tell them why their trainees failed, if they did.

These are some of the major problems encountered by test developers, though by no means all. For the most part, the kinds of test development problems that we encounter in the field of training evaluation are not the same as those encountered in the field of aptitude testing. Thus, we have found the traditional body of academic literature on test development to be poorly suited to our needs. Certainly the basic notions of reliability and validity apply to any test development effort, but in our field, the exotic, sophisticated formulas that fill most books on test development are of little use.

One vital need in the field of training evaluation, it seems to me, is a how-to-do-it manual for test developers—one that responds to the variety of practical constraints and problems that occur in the process of constructing tests for the myriad tasks spanned by some eight or nine hundred Army jobs.

I wish that I had such a manual for you, but I don't. What I do have is intended to be a step, albeit small, in that direction. I have attempted to chart the major action points in the course of developing a test for training evaluation. These steps in performance test development are shown in Figure 1, and I hope that you will find it a useful framework for discussing the problems and practices of test development.

There are two matters of terminology that need clarification. The first has to do with the concept of performance testing. I choose to use this concept (at least today) to designate the test or tests, normally developed and administered by a quality control agency on completion of training for the two explicit purposes of qualifying trainees and evaluating training. This type of testing is to be distinguished from the development and use of tests by trainers for monitoring student progress within and between stages of training. The second is that I use the term *test item* in referring to the evaluation of behavior involved in a single job task, and the term *test* in referring to the aggregate of these items over an entire job or job sector purportedly covered by the training program. I am not asking you to agree with these labels, but to bear them in mind for the moment.

Now let us return to the process of test development as outlined in the figure. I should like to proceed through the 14 steps, and give a brief summary of my thoughts on the "why, what, and how" of each one.

The first three steps on the chart concern assembling information that should routinely be supplied to the test developer. He should only have to verify completeness of the information; and not make judgments about its accuracy. As stated in the first step, test development begins with the objectives for the job or job sector for which people are to be trained. These are sometimes termed job objectives—more often, terminal training objectives. Whatever they are called, they are the master list of specifications derived from the job, and from which *both* training developers and performance test developers, sepatately, begin their work. As test developers, our goal is to develop a performance test item for each and every objective, although this is not to imply that our final test will necessarily encompass all objectives. In addition, each objective should be accompanied by a supporting list of skill and knowledge requirements to be used in later stages of test development.

The information designated in Step 2 should also be available as a matter of course. The relative importance of each objective, as judged in terms of mission capability, represents data that is necessary in making trade-offs later in the test development process.

Step 3 suggests that each objective must be reviewed to make sure it is all there. We know that, in addition to a stated task behavior, an objective should contain stated conditions and standards of performance. If any of the three elements are missing, or if any are unclear to the test developer, he should get together with the task analyst and, as indicated in Step 4, obtain a clear statement of the missing or confusing elements. Performance standards are the most common source of trouble, and if a fair and meaningful pass-fail criterion is to be established for a test item, the developer must have an unequivocal standard of task performance to work from.

In Step 5, test item development really begins. Here, the developer must judge the feasibility of duplicating in a test situation the conditions and behavior called for in the objective. Normally, of course, our view is that well stated objectives are blueprints for testing—in fact, dictating what the test conditions will be. Occasionally, however, we encounter an objective calling for the use of job-relevant equipment, terrain, support personnel, or a time frame that exceeds the resources available to the test agency. In these instances, the developer must carefully weigh the criticality of the objective (from Step 2) against the cost factors before deciding that full realism cannot be afforded, because invariably some degree of relevance is lost as one departs from the test specifications given in the objective.

When it is decided that the conditions of the objective cannot be duplicated in the test situation, a substitute technique must be developed, as indicated in Step 6. This is perhaps the most subtle and challenging aspect of the development process. Here, a developer's inventiveness is often needed in devising a method and conditons for testing that will call for the demonstration of a behavior that is as similar as possible to the behavior stated in the objective. Too often in this situation developers resort to paper-and-pencil tests measuring knowledge of the task, an approach that in most cases can be safely rejected out of hand. In considering simulation options developers have a useful check available in the task's skill and knowledge requirements. The relevance of a proposed test method may be evaluated by checking the number of skill and knowledge components of the task that are called for in the method.

Once a task-relevant method of testing is determined, Step 5 or Step 6, the developer turns his attention to the matter of achieving measurement reliability. In Step 7, he must again look at the objective in terms of repetitions or variations of the behavior implied. In most cases, this will be explicitly given. For a specific skill, such as disassembling a rifle or installing a carburetor, a single demonstration of the behavior is all that is normally called for. On occasion, however, with generalized skills or generalized behaviors, the number of repetitions of the behavior may or may not be clearly stated in

the objective. An objective specifying that something will be done correctly 9 out of 10 times creates no problem for the test item developer, as 10 repetitions are required. On the other hand, the standard may be phrased in terms of correct performance on 90% of the trials. Here a decision must be reached on an appropriate number of repetitions of the performance to ask for in the test item. More generally, the important consideration in Step 7 is whether a large enough sample of trainee performance is being required so that success or failure does not result largely from chance. Here, again, the test developer must make some trade-off between time or cost factors and reliability of the measured behavior.

Step 8 pertains to another aspect of test reliability—the standardization of the conditions under which a test item is administered. Here, the important factors are the instructions and environmental conditions under which the test item is given. Instructions should be identical for everyone. They should be clearly and simply stated, leaving nothing to the interpretation or misinterpretation of the trainees taking the test. Things such as the method of scoring and whether speed or accuracy is important should be stressed in the instructions. Also, conditions pertaining to test supplies and environmental factors should be constant for all personnel. Items of equipment worked with or on during testing should be restored to their pretest condition if they are used by successive trainees. Similarly, environmental factors such as visibility, temperature, attitude of the tester, time of day, and the like, must be stabilized.

In Step 9, a final aspect of measurement reliability is considered. Here procedures for translating an observed trainee performance into a pass-fail score must be developed. Provision for this type of scoring should be structured so that only the more reliable human skills are used. That is, the scoring activity should be reduced to one of matching or comparing the test item response with some model of the acceptable response. If the model response on a test of rifle marksmanship is defined as a hole in the bullseye, then the scorer has a relatively easy task in judging the acceptability of the response made by the rifleman. Unfortunately, responses for many test items cannot be judged in this "either/or" fashion, but require a "more-or-less" type of judgment. In these cases, the developer should always strive to break down the model response into elements so that comparative judgments can be made more easily by the scorer. This may often entail preparing a checklist of the necessary components or features of the model response.

In Step 10, a supplementary scoring procedure is developed for use in diagnosing reasons for trainee failure on the test item. Pass-fail scoring is sufficient in meeting the primary mission of quality control, which is the certification of trainee job readiness. However, the secondary mission, that of training program evaluation, is best accomplished by providing the trainers not only with the incidence of pass and failure for an objective but also feedback on why trainees failed. One way to obtain this data is through a checklist developed from the skill and knowledge requirements of the task to be used by the tester in recording why the trainee failed a test item. When accumulated over a number of test item administrations, this diagnostic information will normally provide a stable picture of the reasons for failure that trainers may then use to selectively revise and strengthen their program.

In Step 11, the test developer simply brings together the products of previous steps and formats the final test item. Detailed instructions to the tester covering test materials, equipment, procedures, precautions, and so forth, are spelled out. The directions to be read to the trainee by the tester, and the scoring procedure should also be written out.

The final three steps in the figure pertain to assembly and administration of the final form of the test. In Step 12, a decision is made on whether time permits testing on all objectives—that is, administration of all test items. If it is not feasible to do so, an appropriate sample of test items has to be selected (Step 13). As indicated in this step, the main criterion for sampling should derive from criticality ratings of the objectives. An

exact procedure for doing this will depend upon the categories originally used for reporting criticality. Generally, the developer would first include all "essential" or highly critical items, and then sample from the remaining. Wherever sampling is necessary, the usual practice is to vary the sample from one administration to the next so that all test items are used sooner or later. Variations in the sample should not be systematic in the sense that trainers or trainees can anticipate what items are going to appear.

In Step 14, final guidance for test administration is prepared. Training for testers may have to be developed; lists of equipment and materials prepared; and scheduling worked out. If testing is to be done individually, it is usually a good idea to prescribe a "county fair" layout of test stations. This serves purposes of economy, as well as permitting test items to be administered in varying order. In addition, security precautions must be specified to ensure, for example, that one trainee cannot benefit by observing another's performance, or that trainees do not talk among themselves during test administration.

Consideration of these action points, step by step, constitutes a framework for performance test development.

Appendix C

Guidelines for the Evaluation of
Applied Performance Test Materials and Procedures

GUIDELINES FOR THE EVALUATION OF APPLIED PERFORMANCE TEST MATERIALS AND PROCEDURES

General guidelines are proposed for the evaluation of Applied Performance Testing situations. Since Applied Performance Testing is conducted within a wide range of situations, these guidelines should be applied judiciously to individual instances. There may well be times that very good Applied Performance Tests do not conform to some of the guidelines which are included. In general, however, good Applied Performance Tests are expected to demonstrate the qualities represented by the proposed guidelines.

> Definition: "Applied Performance Testing, for purposes of the Clearinghouse for Applied Performance Testing (CAPT) project, is defined as the measurement of performance of some task significant to a student's life outside the school and/or to adult life. Such a task is valued as output for public schools. The testing device must allow for measurement of the task in an actual or simulated performance setting."

A SUMMARY OUTLINE OF GUIDELINES FOR THE EVALUATION OF APPLIED PERFORMANCE TEST MATERIALS AND PROCEDURES

1.0 Test Background

- 1.1 Purpose of the Test
- 1.2 Test Content
- 1.3 Task or Job Analysis
- 1.4 Pilot Testing and Validation

2.0 Characteristics of a Good Test

- 2.1 Validity
- 2.2 Reliability
- 2.3 Adequacy
- 2.4 Objectivity and Standardization
- 2.5 Comparability
- 2.6 Efficiency/Practicality
- 2.7 Balance
- 2.8 Difficult/Discrimination
- 2.9 Fairness
- 2.10 Speededness
- 2.11 Format
- 2.12 Relevance

3.0 Test Administration and Reporting

- 3.1 Instructions to the Examiner
- 3.2 Instructions to the Examinee
- 3.3 Scoring

1.0 Test Background

1.1 Purpose of the Test

1.1.1 The purpose of the test should be explicitly stated to aid understanding by examinees, users of test results, and those administering and interpreting the test.

1.1.2 The construction of the test should reflect the purpose for which the test is to be used.

1.1.2.1 Selection tests need to discriminate well around cut-off points but would require only limited discriminating power if the ends of the distribution standards for selection need to be well defined.

1.1.2.2 Certification tests need only discriminate between individuals who have and those who do not have certain well defined competencies.

1.1.2.3 Applied performance tests, of diagnostic nature, need to cover a limited scope but in much greater detail. Such tests should be designed to yield scores on separate parts. The range of item difficulty and individual discriminating power is less important.

1.1.2.4 Applied performance tests for classification of performers need to have a sufficient range of item difficulty and individual discriminating power to differentiate

individuals on a continuum of expected competencies. Such tests are expected to yield a single score and are expected to be more general in nature than tests for comprehensive testing and description of competency levels of specific individuals.

1.2 Test Content

1.2.1 Tests should measure the performance of some task thought to be significant to a student's adult life or life outside of school. (Example of tasks: (1) read and comprehend the front page of a newspaper, (2) make change (money), (3) read and follow directions on a medicine bottle, (4) complete an application for a job.)

1.2.2 Tests should provide for the measurement of the task in an actual or simulated performance or job setting.

1.2.3 Tests should measure useful abilities of a practical nature that contribute to success in life or success in some aspect of the world's work.

1.2.4 A paper-and-pencil test can be considered the most appropriate applied performance test when the test response is identical with the behavior about which information is desired. For example, a test in accounting or shorthand would have to use the paper-and-pencil format.

1.2.5 Test content should provide reasonable items which sample/depict those behaviors in extracurricular

or adult life activities that are consistent with the social and cultural contexts in which the activities occur.

1.3 Task of Job Analysis

1.3.1 A task or job analysis should be referenced when developing applied performance tests for testing in complex job situations. A job analysis can include information about job training, responsibility, job knowledge, dexterity and accuracy, and equipment, materials, and supplies. Also, information is needed on examples of situational factors that are commonly associated with and which may affect task performance. For example, the condition of a patient involved is a highly significant variable in describing a task.

1.3.2 The test should show evidence of the representativeness and criticality of tasks and sub-tasks to be measured. Sub-tasks should be capable of impact on overall task fulfillment.

1.3.3 If all the elements of a job are not measurable because of constraints on time or resources, the sample performance elements to be observed should be identifiable as the most critical or crucial aspects of a job. Critical representativeness of the performance elements to be sampled should be apparent.

1.3.4 There should be information on the relatedness of what is being measured to the kind of information

needed such as accuracy in job performance, speed in task completion, and dexterity in the use of tools.

1.3.5 Should the "state-of-the-art" deter measurement of the most critical elements of a job, some references should be available to the inherent problems of measurement in relation to the critical elements.

1.3.6 When it is difficult to identify actions or behaviors that constitute successful performances, there should be information on the relatedness of such behaviors to profiles of persons who are considered competent or skilled. Profiles can be represented in the form of task performance checklists.

1.4 Pilot Testing and Validation

1.4.1 Evidence should exist to show that the measure has been pilot tested. Particular attention should be paid to, among others, the following criteria:

(a) Directions are clear and unambiguous.

(b) Rating/scoring procedures are feasible, accurate and objective.

(c) Time limits, if any, are reasonable and consistent with the objectives.

1.4.2 Pilot testing should be conducted to identify the test items which discriminate well between persons who are competent at the task and those who are not.

1.4.3 In developing an occupational competency test, there should be evidence that experts in the field scored perfect or near perfect scores on the pilot test in terms of product.

1.4.4 As evidence of specificity of measures to be obtained, a near chance score or better.

1.4.5 Tests should be independently reviewed by (1) employers and practicioners who will eventually judge the competence of a performer and (2) panels of reviewers who should be carefully chosen and their qualifactions fully documented. Tests should be reviewed for relevance, clarity, feasibility, and appropriateness of purpose.

1.4.6 When testing non-verbal skills, pilot testing should adequately demonstrate that students with limited verbal skills of English-speaking competencies can fully understand what is expected of them.

2.0 Characteristics of a Good Test

2.1 Validity: How well does the test measure what it purports to measure?

2.1.1 Content Validity: Does the test require a demonstration of competencies representative of the knowledge and skills required for the task or activity being measured?

2.1.2 Concurrent Validity: Is there evidence of substantial correlation between the test, especially one involving simulation or one shortened to include only selected tasks, and a reliable and valid independent criterion of performance?

2.1.3 Predictive Validity: If the test is used for prediction or selection, is there subsequent evidence that the test served as a good predictor of competency?

2.1.4 Claims of validity should be appropriately documented. Such claims include, for example, correlations between measures of performance on test items representing a domain of well-defined test situations and one or more measures or performance "on the job" or in the "real" world.

2.1.5 Whether the actual test situation should differ from the exact situation in which the skills would be applied depends on the nature of the task. There are instances in which the domain of task conditions is so large that training must focus on a generalizable

principle rather than on teaching a response to every possible set of conditions. It is assumed that the student should be prepared for all possible test situations and should not be totally surprised by what he/she encounters on a test.

2.1.6 When direct application of skills or competence is not observable/measurable, testing may have to be limited to provisions of indirect, inferential evidence of proficiency, such as possession of the essential aptitudes and prerequisite skills.

2.1.7 The task performance to be observed should be so highly structured that variation in results can be attributed to different levels of competency in students' performance of a task and not to be extraneous factors related to the measurement instrument of technique.

2.1.8 When testing in a simulated setting, the simulation should be close enough to reality to be satisfactory for a given purpose.

2.2 Reliability: How well does the test measure what it is intended to measure? Are test scores consistent and dependable? Have those sources of variation which are attributable to chance been eliminated or controlled as far as practical?

2.2.1 Documentation of reliability coefficients (parallel form, test-retest, split-half, or internal consistency), when used appropriately, should provide for measures of reliability based on the variance in the

proportions or frequencies of correct, incorrect, and not attempted responses across equivalent sample test items. (Caution: the split-half and internal consistency measures are not appropriate in cases of performance tasks with sections that do not involve the same skills or that include skills which are discrete and not necessarily correlated. If parallel forms are to be used, care must be taken to check on the effect of altering variables on the two similar test forms. When it is practicable to have stability coefficients for measures of an individual person's competencies, repeated testing should be conducted.

2.2.2 When having more than one form of the test is practicable, parallel form reliability is the preferred form of reliability.

2.2.3 When scoring procedures require judgment, the reliability of such procedures should be documented by showing the degree to which several independent judges score performance in a like manner.

2.3 Adequacy

2.3.1 The test should be of sufficient length and scope to sample appropriately and faithfully the behavior it is designed to measure.

2.3.2 Tests shou provide, when feasible in terms of time and resources, a sufficient number of varying trials to verify the results as good measures rather than an accident of chance.

2.3.3 When practicable, it is highly desirable to have parallel forms of the test.

2.4 Objectivity and Standardization: The test should be construed in such a way as to control or eliminate the influence of random factors, personal opinions, and unreliable subjective judgments on the final results.

2.4.1 It should be possible to present the task in virtually the same manner to each examinee.

2.4.2 For standardization of instructions to examinees, a tape recording of all instructions should be considered as well as an accompanying written text for the examinee.

2.4.3 Any equipment which is used should be subject to standardization regarding its technical features and proposed use.

2.4.4 Unintentional clues which are included in the instructions or in other test items should be meticulously checked for and eliminated.

2.4.5 When a performance instrument has been tested for reliability in a specific testing situation, one must be careful about using that instrument in other similar situations in which testing conditions differ even slightly, for even minor variations can place test results in question. For example, a test designed to measure ability to drive in metropolitan areas may not be appropriate for measurement of rural driving ability. The driving skills are basically unchanged, only the situation has been varied.

2.4.6 . Acceptable procedures for controlling objectivity are: (a) percentage of agreement among independent observers, (b) correlation among independent observers, and (c) Guilford's analysis of variance among raters and ratees.

2.5 Comparability

2.5.1 The results of the individual examinee's test, as necessary, should be subject to meaningful and objective interpretation through comparison with other test scores obtained by the examinee, or comparison with predetermined criteria or other standards set for the examinee.

2.5.2 If norms have been established for the test, explicit and complete descriptive information about the norming population and the procedures used should accompany the test.

2.5.3 Whether the test score interpretation is norm-referenced or domain-referenced should be determinable and consistent with the purpose and type of information desired.

2.6 Efficiency/Practicality

2.6.1 The results of the test should be worth the amount of time, effort and money required of both examiner and examinee to obtain those results.

2.6.2 Results should be critically reviewed to determine if a less complex mode of testing could have obtained equally useful information. For instance, would a paper-and-pencil test have been just as good? Would

a simulation test setting have been just as appropriate as a real job setting? Acceptability of the paper-and-pencil test or simulation test setting should, however, be based on identifiable criteria such as the cost of an error in a real life situation in terms of time, people, and resources.

2.6.3 Tests which can be administered just as well on a group basis should be administered in this manner. For the sake of efficiency, some validity may have to be sacrificed. The decision must be based on comparison of the relative advantages of each approach.

2.6.4 Supplies, tools, and equipment to be used should be held to a minimum, but should be adequate to ensure a realistic measurement of the task.

2.6.5 When "process" (or work procedure) is not both necessary and sufficient to completing the task "product" (outcome), the test should include provisions for measuring the task product.

2.6.6 For tasks of long duration, in which many people are to be tested, a sampling scheme of people at various points in time should increase efficiency when only group data are needed.

2.6.7 The performance to be observed should involve as little repetition of identical procedures as possible in any one testing. A single item, if well constructed, can be highly reliable.

2.6.8 Some correlation should be evident between developmental costs and projected relevance, accuracy, and use of test items over time. Would the items be of use long enough to justify developmental costs? Would the items be free of errors that could result in or contribute to fatality in critical testing situations--for instance, in the medical field?

2.6.9 When the testing situation involves human stress, any simulation effort should be considered from a practical, legal, and ethical standpoint.

2.7 Balance

2.7.1 The testing time and the importance or significance attributed to the results of each task should generally correspond to the instructional importance or priority of the task itself.

2.8 Difficulty/Discrimination

2.8.1 The difficulty level of the task and its complexity should be appropriate for the maturity of the examinees.

2.8.2 Tests which are used for classification and diagnosis should include a sufficient number of "easy" items in addition to "hard" items to permit meaningful analyses of examinees' strengths and weaknesses.

2.8.3 Irrelevant sources of difficulty (e.g., inappropriate reading level, vague directions, unclear illustrations, poor quality of test materials) should be eliminated to the extent possible.

2.8.4 Except in domain-referenced testing, tasks which do not discriminate (e.g., "everybody can do them" or "nobody can do them") should be reevaluated when considered for further use. Items which do not discriminate add little to the quality of a test except in domain-referenced testing.

2.8.5 A test designed to make very fine measurement distinctions should contain multiple items of a sufficient number to permit general conclusions.

2.9 Fairness

2.9.1 The verbal factor in tests (reading, writing, speaking) should be minimized in the testing of specific performances except in those cases which require oral/written/reading communication skills.

2.9.2 Tests should be carefully checked for irrelevant sexual bias or content bias against any ethnic, social or geographic group. Such a check could be based on reviews and empirical studies.

2.9.3 Tests should measure those skills or areas of knowledge which are based upon instructional objectives considered valid for the examinees involved.

2.10 Speededness

2.10.1 The time allowed examinees should be appropriate for the length of the test.

2.10.2 In those cases where the speed with which a person works is <u>not</u> important, the test should allow enough time so that all examinees have time to finish the test.

2.10.3 In performance test situations where speed is an important indicator of competency, the number of items completed in a set time or required length of time should be set as a measure of successful performance. Examinees should be told whether time is a factor in scoring.

2.11 Format.

2.11.1 To help examinees develop positive attitudes toward testing, and to sustain examinees' interest during the testing process, motivational factors such as (a) novelty of stimulus, (b) attractiveness of stimulus, and (c) action-orientedness should be incorporated in the test situation.

2.11.2 Procedures for testing in the affective domain should be based on an unobtrusive and ethical test method. Motivational or attitudinal tasks cannot with certainty be validly tested by conventional means; tasks may be better tested covertly through ongoing observation or structured observation during performance of some task that is ostensibly being tested.

2.11.3 When attitudes must be observed and measured in a contrived obtrusive setting, the stimulus should include provisions for helping examinees respond meaningfully and realistically. For example, in measuring value judgments, an examiner might show examinees a short film of an emotional situation and ask examinees to evaluate the situation; in the process of evaluation, examinees' value judgments would be more

meaningfully elicited than through traditional paper-and-pencil inventories.

2.11.4 Generally speaking, if a task's natural sequence is not critically disturbed, it is desirable to have test items or tasks progress from simple to difficult.

2.12 Relevance

2.12.1 Tests in use over time should be periodically re-evaluated whenever instructional objectives or performance requirements are changed to any considerable extent. Test items should reflect current, updated, material.

3.0 Test Administration and Reporting

3.1 Instructions to the Examiner

3.1.1 The procedures to be followed by the examiner should be clearly specified.

3.1.2 All instructions to examiners should be as simple as possible.

3.1.3 The equipment, facilities, or other materials to be used should be clearly specified for the examiners.

3.1.4 Detailed guidance should be given the examiner as to the type and limits of assistance (oral or other) that may be given examinees.

3.1.5 Detailed guidance should be given the examiner covering the physical layout and the management of facilities, and the testing time necessary to ensure that examinees are tested fairly, efficiently and without jeopardizing test integrity. (This is most important in conducting large-scale, concurrent testing of individuals at multiple test stations.)

3.1.6 Any potential hazards or safety precautions to be taken should be pointed out to examiners.

3.1.7 Equipment and materials used by successive examiners should be restored to pre-test condition for each student.

3.1.8 Test users should be advised in understandable terms of the limits and constraints, applicability, and interpretations of test results.

3.2 Instructions to the Examinees

3.2.1 The purpose of the test should be explained.

3.2.2 Time limits, if any, should be explained.

3.2.3 The equipment, facilities and other materials which are available should be specified for examinees.

3.2.4 Any safety precautions or potential hazards should be noted for the examinees.

3.2.5 The process of answering items or demonstrating competencies as well as the method of scoring should be carefully prescribed.

3.2.6 Examinees should understand how much freedom they have in demonstrating competency and whether they are subject to penalty for guessing.

3.2.7 All instructions to examinees should be as simple as possible.

3.2.8 A procedure should be included to ensure that the examinees know what they are expected to do. Responding to a sample question is an example of such a procedure.

3.2.9 When it is expected that the test format is new to the examinees, they should be given some practice, in advance, using that format.

3.3 Scoring

3.3.1 Scoring procedures should be standardized and objective.

3.3.2 When completeness of performance is to be observed, performance at each checkpoint should be scored as "passed" or "failed;" each test item should be unambiguously scorable as either correct, incorrect, or not attempted.

3.3.3 When rating scales are used, rating categories should be carefully defined with specific examples given as a standard of comparison for each category; scale points should be sufficiently discriminating.

3.3.4 When possible, multiple judges who are well trained are preferable to a single judge. There may be occasions when one well-trained judge is preferred, if the quality of other prospective judges' training is questionable.

3.3.5 Interjudge reliability should be established and documented with all scoring procedures.

3.3.6 If there is more than one judge, each should make judgments independently, with subsequent negotiation to reach consensus on the rating to be assigned.

3.3.7 It should be determined in advance whether the "process" or the "product" of the task will be more important in scoring. (In many cases some combination of the two will determine the score.)

3.3.8 Generally, both the "quality" of the work and the performance "time" considered in scoring are dictated by the task; thus, standards for scoring should be documented.

3.3.9 The scoring method for the test should be consistent with the purpose of the test. For example, if the test is being used to determine examinees' progress over time, can the score information be appropriately used to show change in performance over time. Can

the score information be used for trend analysis over time or with different groups?

3.3.10 Scoring keys and procedures should be pilot-tested and checked for feasibility, clarity and appropriateness.

3.3.11 Maximum use should be made of scoring aids, such as templates, to further the objectivity of scoring.

3.3.12 Detailed instructions on how to score the examinees and provisions for practice scoring trials should be provided.

3.3.13 The number of tasks to be scored or rated should be sufficiently moderate that the rater(s) can score accurately and reliably.

3.3.14 Specific scoring guidelines, criteria and required examinee qualifications for scoring on the basis of direct observation should be specified.

3.3.15 The scoring of trivial tasks should be avoided.

3.3.16 Whenever possible, scoring should be done without examinee identification to minimize biases and inconsistencies.

3.3.17 To the extent possible the scoring activity should be reduced to one of comparing the test item response with some model of the acceptable response. If a response cannot easily be judged in a "yes/no" fashion, but requires a "more-or-less" judgment, the model response should include enough examples to permit reliable comparative judgments.

3.3.18 The feasibility of making audio or video recordings of task performance should be considered, since this permits a more accurate scoring procedure. This is particularly useful when the task process is transient or does not result in a product that can be examined at leisure by the examiner.