DOCUMENT RESUME

ED 124 585 TM 005 341

AUTHOR Sanders, James R.
TITLE Measurement Problems and Issues Related to Applied Performance Testing.
PUB DATE [Apr 76]
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Elementary Secondary Education; *Measurement; *Performance Tests; *Problems; *Test Reliability; *Test Validity
IDENTIFIERS *Applied Performance Testing

ABSTRACT

Applied Performance Tests (APT) are defined as instruments designed to measure performance in an actual or simulated setting. They require at least a close approximation of the setting (if not the actual setting) to which the performance is expected to be transferred. This paper outlines measurement problems and issues that are unique to APT. It is argued that the problems and issues that are widely discussed for criterion referenced tests are also appropriate to APT. A brief history of APT is given. A listing of reliability and validity problems unique to APT is presented and discussed. Two additional measurement problem areas in APT are their objectivity and the generalizability of their results. Other measurement related considerations that may be regarded as problems in APT include cost, difficulty of application and development, and unavailability of norms for test interpretation. Finally, research and development steps to address the shortcomings of APT in elementary and secondary education are listed. (RC)

MEASUREMENT PROBLEMS AND ISSUES RELATED TO APPLIED PERFORMANCE TESTING[1]

James R. Sanders
Western Michigan University

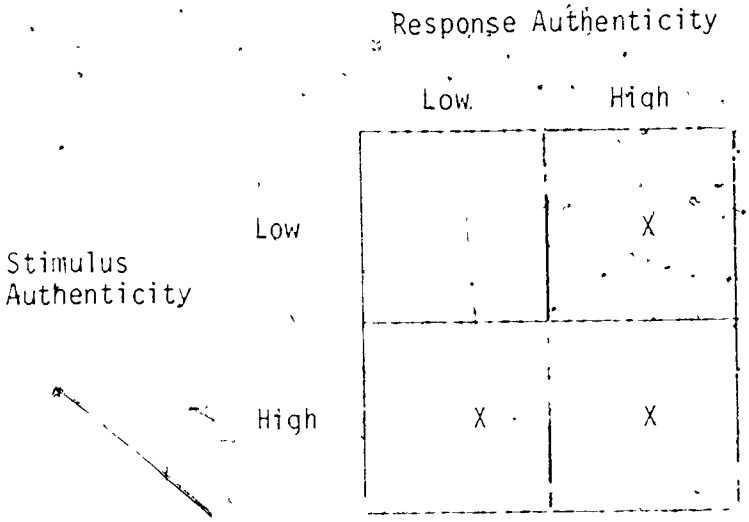
Applied Performance Tests (APT) have been defined by Sachse and Sanders (1975) as "instruments designed to measure performance in an actual or simulated setting." They are measurement devices that require at least a close approximation of the setting (if not the actual setting) to which the performance is expected to be transferred.

The purpose of this paper is to outline measurement problems and issues that are unique to APT. I would argue that the measurement problems and issues that are widely discussed for criterion referenced tests (e.g., Harris, Alkin, and Popham, 1974) are also applicable to APT. In order to limit this discussion, and because there are many fine discussions of problems and issues that APT holds in common with other tests, I will concentrate on some of the more salient measurement concerns that are unique to APT.

The uniqueness of APT is found in the high degree of realism built into the test. (Realism, fidelity, and authenticity are used interchangeably in describing the degree to which these tests reflect real life situations that require the behaviors being measured, following Sachse and Sanders [1975]). Both exercise stimuli and response modes can serve as focal points for applied performance test designation. Both the stimulus and response can either have high or low fidelity. If either have high authenticity, the

[1]Comments prepared for a Symposium on Applied Performance Testing: Research and Development Perspectives. Held at the annual meeting of the American Educational Research Association, San Francisco, California, April 1976.

instruments is generally classified as APT. A figure reproduced from Sachse and Sanders (1975) depicts the instruments that may be classified as APT where X's denote APT situations:



Examples of tests that would fall into these categories were provided by Sachse and Sanders (1975) and are not reproduced here.

Tests may be classified in many different ways. For example, we might classify them as measures of cognitive, affective, or psychomotor behaviors. Or, we might classify them in terms of maximum versus typical performance, following Cronbach (1970). Attempts to classify APT using these categories usually fail, however, indicating theoretical inadequacy in such classification schemes. The reasons for such failure fall from the nature of the performance being observed using APT. The performance might be an emotional response to some stimulus, or a psychomotor performance. It usually involves knowledge about appropriate responses. In fact, the performances that are typically recorded using APT involve a complex combination of each type of behavior. In this sense,

then, APT may be thought of as molar instrumentation (not in the dental sense) rather than instrumentation used to measure molecular or elemental behaviors. But even this distinction breaks down in that molecular responses, if they have high authenticity, could be measured by APT. The psychological theory underlying the development and use of APT is not well developed and leads us to problems of definition, classification, and interpretation with APT. Although some would argue that this is not a measurement problem or issue, it is important to note.

Some History

Historically, APT has been a mainstay for military and occupational testing for years. Reviews by Fitzpatrick and Morrison (1971) and Panitz and Olivo (1970), added to the volume edited by Glaser (1962) provide a fine overview of the development and use of APT. Professional occupations, especially the medical arts, and business and industry have a shorter, but productive, history. The field of public elementary and secondary education has little history in the use of APT, with interest just now developing in the areas of teacher evaluation, measurement of student achievement, and teacher and administrator training. The forms of APT that have been developed and used in the military, in occupational examination agencies, in medical centers, and in business and industry include the following:

Military APT	Occupational APT	Medical APT	Business & Industry APT
simulation gaming situational tests	work products on-the-job process observation	simulation situational tests including problem solving tests	simulation gaming situational tests including in-basket tests

All forms of APT have been used by each, no doubt, but the forms listed in each column appear to be those that have received the most emphasis.

Considerable interest in forms of APT for use in elementary and secondary achievement testing has appeared recently. In a search for users of APT in public school content areas we found considerable variance by content areas. Reading, mathematics, and physical education at the secondary level included frequent use of APT. This was considerably less true at the elementary level. Content areas that appeared to be void of APT materials included the social sciences (history, civics, psychology, philosophy, and economics), the arts (drama, literature, and art and music forms), the physical sciences (geology, geography, biology, chemistry, and physics) and, surprisingly, the area of foreign language study. However, the fact that formalized, widely available applied performance tests were not found in many public school content areas does not mean that APT is not used. Rather, performance measurement that does occur usually takes place in an informal manner. The potential for developing standard APT materials for public school use of the focus listed above is great. It remains a challenge to those who develop measurement devices to provide APT for use by educational practitioners. An examination of measurement problems and issues unique to APT should provide some guidance to this effort.

Measurement Problems and Issues Unique to APT

By identifying realism of stimulus and/or response as the unique characteristic of APT, we have narrowed our discussion to measurement problems and issues created by this requirement. At first glance it is tempting to conclude that there are few measurement problems and issues that are unique to APT, but further investigation suggests otherwise.

Consider, first, the reliability of APT. Certainly we have the tools to calculate reliabilities, depending on the form of APT being developed:

1. For simulation, gaming, and situational tests where mechanical or paper and pencil responses are used, the KR-20, or, under specific conditions, the alternative ways we have for calculating reliability on paper and pencil tests are appropriate.
2. For rating or ranking work products, interjudge reliability, the coefficient of concordence, or nonparametric tests we have for ordinal data are sufficient.
3. For process observation, the same techniques we have developed for determining the reliability of the many classroom observation schedules that exist are appropriate.

What problems can exist? A listing of reliability problems that are unique to APT includes:

1. Control over the testing environment. As the realism of APT is increased, a greater number of extraneous variables are introduced into the test. Irrelevant, often random, cues on the stimulus presentation will certainly affect the examinee's response. Obstructions to the examinee in giving the response he would normally give will also affect his performance. Thus, testing under real conditions will frequently lead to measurements with low reliability.
2. Number of times one examinee may be tested. It has been suggested (e.g., Gagne, 1962) that repeated measurements on an individual, where several tasks of the same type are given, may serve to increase the reliability of APT. However,

considering the cost of APT (time, facilities, risk, logistics), often only one trial is possible. The reliability of this one trial is usually low.

3. Problems with instrumentation. When hardware is being used to record examinee responses, reliability is usually not a problem. However, when human recorders are used, observer variation can adversely affect the reliability of the APT. Webb et al (1967) have addressed this problem in detail.

4. Other variations in testing conditions. Conditions in the testing environment, as noted earlier, can affect the reliability of the measure. The standardization of APT administration can improve the reliability of the tests, but can also remove realism from the testing situation. Standardization of directions and administration time are two concerns that should be addressed. They can also affect the validity of the test. Added to this problem are variations due to time of day, month, or year and psychological and physical state of the examinee. These, too, affect the reliability of the measurement, although much the same could be said about other tests as well.

Another consideration is the validity of APT. The criterion validity of APT is important if such tests are to be used in drawing conclusions about one's ability to perform certain valued tasks. Smith (1975) provides a nice discussion of the criterion problem and his discussion

certainly applies to APT. Basically, the validity problems and issues related to APT include:

1. Identification of the ultimate criterion task and demonstrating, empirically, a relationship between performance on an APT and performance on the criterion task. This is an easy-sounding undertaking that has proven to be quite difficult. Task analyses in the military, various occupations, and in the medical arts have proven to be productive and form a basis for many APT materials (e.g., Osborn, 1975, and the many available HumRRO publications). This process has proven to be much more difficult in developing APT materials for public school use, especially when affective performance is of interest. The intervening experiences of people throughout their school years and between the time they receive their secondary diplomas and are called upon to perform certain valued tasks are powerful. This problem is an important one to be dealt with in developing APT materials for public school use that do tell us something about ultimate criterion performance.
2. Control over the testing environment. The closer to reality APT moves, the higher the criterion validity of the measurement. However, as I noted earlier, reliability is usually lower under realistic conditions and, as we know, the reliability of the test does place limits on its criterion validity. As we gain control over the testing environment, the criterion

validity is usually lowered, although the reliability of the APT is increased. This trade-off presents a tough problem to those wishing to develop APT for public school use. There is no good answer to the questions of how much control is appropriate or how far APT can be removed from reality and still have criterion validity.

3. Identifying effective stimuli in APT and determining valid scores. This problem is related to the first reliability problem discussed earlier. It is difficult to standardize test stimuli in many real-life situations and, hence, two different examinees may actually be performing different tasks within the same APT. For example, one examinee may receive a high score on an APT only because he undertook the easy elements of the total task performance while leaving the difficult parts go. Another examinee may receive a low score because he undertook the tough parts first and failed. Standardization of testing conditions and scoring procedures presents a difficult measurement problem to those who wish to develop APT for public school use.

Two additional measurement problem areas in APT are the objectivity of such measures and the generalizability of their results. When hardware is being used to record the performance of an examinee, objectivity presents little problem. Certainly airline pilot simulators that mechanically record the responses of examinees provide little room to doubt the objectivity of

recorded scores. However, when observations, ratings or rankings of products, or other human recording devices are used, the objectivity of the measurement is a problem worthy of consideration and safeguards against bias need to be built into the data collection and scoring procedure.

Many of the comments I made earlier about reliability and validity problems of APT relate to the problem of generalizability of results. Standardization of testing conditions, criterion validity, intervention of extraneous variables into either stimulus conditions or responses, and scoring procedures all help determine whether an examinee's reported performance can be generalized to other settings, other persons, or other times. Because these are problem areas with APT, one has to include the generalizability of APT scores as a problem also. From generalizability theory we all know that no one observation of performance can be considered representative of the person's ability to perform. Measurement limitations of APT limit generalizability of scores even further than the limitations of commercially available achievement tests that schools now use.

A listing of other measurement related considerations that may be regarded as problems in APT include cost, difficulty of application and development, and unavailability of norms for test interpretation. I suspect others in the Symposium will be discussing these concerns so I leave it to them to elaborate.

Implications for Research and Development

Is APT to be avoided in elementary and secondary education because of these shortcomings? I don't believe it should. In fact, I believe there

is a great amount of yet unrealized potential in APT for public school use. APT, to be sure, is just one small part of the entire testing spectrum used in our schools. It is not a panacea for testing problems in education nor is it a replacement for the many highly developed technical tools now used. It is a way to get information about the performance of people on certain valued tasks. At present, this limited type of testing is underdeveloped and underused in education and I believe it is productive to examine ways that we can address some of the shortcomings that I have mentioned.

I would suggest the following research and development steps to address these shortcomings:

1. Curriculum and measurement specialists need to work together in identifying tasks that are important in their own right and those that are associated with valued task performance in later life. The focus of this inquiry should be on identifying those tasks that are within the scope of the public school curriculum.
2. Curriculum and measurement specialists need to work through a national association in task forces or funded projects to develop standard APT's that can be made available to schools nationwide. Technical manuals, developed to meet the AERA/APA/NCME Standards for Educational and Psychological Tests, should be produced for these tests. I would expect the measurement problems I have discussed to be addressed further by these projects.

3. Task analysis studies of valued adult performances need to be undertaken and the results linked to public school curriculum. It is important that the elemental tasks for later task performance are systematically covered in the school curriculum.
4. Task analysis studies of valued performance expected when students exit the public schools need to be undertaken and the results linked to the K-12 curriculum so, again, instruction on the elemental tasks is not left to chance.
5. Although the criterion validity of an APT is its most important characteristic, there is a need to examine methods of controlling testing conditions in order to improve the reliability of such measures, while at the same time maintaining high criterion validity.
6. There is a need to systematically study confounding factors that affect APT performance for each form of APT. A taxonomic description of such factors would lead us a long way toward improving the quality of APT materials.
7. There is a need to develop a theoretical foundation for APT. Ways of classifying APT materials do not exist, undoubtedly because of a lack of theoretical structure. Furthermore, it is unclear what different forms of APT measure (i.e., simulations, games, situational tests, process observations, work products). If they measure different constructs, or the same construct, but at different levels

of complexity, a theory should reflect this knowledge. Research into the factorial complexity of APT forms would also contribute to theory development.

8. There is a need for creative development of new forms of APT that may alleviate some of the measurement shortcomings that have been discussed. Educational measurement specialists funded to explore such creative alternatives would contribute new knowledge that would have immediate use for public school testing.

Applied Performance Testing has great appeal for measuring task performance in the public schools. There is much work to be done to refine the concept and improve on our techniques. I believe the effort is worthwhile and expect to see comparatively great advances in APT in the near future.

REFERENCES

Cronbach, Lee J. Essentials of Psychological Testing. New York: Harper and Row, 1970.

Fitzpatrick, Robert and Morrison, E. J. Performance and Product Evaluation. In R. L. Thorndike (Ed.) Educational Measurement (2nd Edition). Washington, D.C.: American Council on Education, 1971.

Gagne, Robert M. Simulations. In Robert Glaser (Ed.). Training Research and Education. Pittsburgh: University of Pittsburgh Press, 1962.

Glaser, Robert. (Ed.) Training Research and Education. Pittsburgh: University of Pittsburgh Press, 1962.

Harris, Chester W., Alkin, Marvin C., and Popham, W. James. Problems in Criterion-Referenced Measurement. Center for the Study of Evaluation Monograph Series in Evaluation, No. 3. Los Angeles: UCLA Center for the Study of Evaluation, 1974.

Osborn, William A. Developing Performance Tests for Training Evaluation. In James R. Sanders and Thomas P. Sachse (Eds.) Problems and Potentials of Applied Performance Testing. Portland, Oregon: Northwest Regional Educational Laboratory, 1975.

Panitz, Adolf and Olivo, C. Thomas. The State of the Art of Occupational Competency Testing. New Brunswick, N.J.: Department of Vocational Technical Education, Rutgers University, 1970.

Sachse, Thomas P. and Sanders, James R. A Look at Applied Performance Testing in Education. Portland, Oregon: Northwest Regional Educational Laboratory, 1975.

Smith, N. L. The Study of Criteria in Educational Evaluation. Unpublished Doctoral Dissertation. Champaign: University of Illinois, 1975.

Webb, E., Campbell, D. T., Schwartz, R. D., and Sechrest, L. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally, 1966.