DOCUMENT RESUME

ED 040 257 UD 010 226

AUTHOR Severson, Roger A.
TITLE Problems in the Practical Fstablishment of Predictive Measures in the Schools: Part Two.
PUB DATE 4 Mar 70
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, Minn., March 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.55
DESCRIPTORS Academic Achievement, *Educational Diagnosis, *Intelligence Tests, Learning Difficulties, Longitudinal Studies, *Predictive Ability (Testing), *Predictive Measurement, Preschool Children, *Reliability, Testing
IDENTIFIERS Bender Gestalt Test, Illinois Test of Psycholinguistic Ability, Stanford Binet IQ Test, Wechsler Intelligence Scale For Children

ABSTRACT

This report discusses the practical problems encountered in a longitudinal study now in its fourth year, where the focus has been the early identification of later learning disorders. The general goal was the identification of tests with the characteristics of high reliability, low cost, short time to administer, low demand on scoring sophistication, and which could be given in group form by the classroom teacher. A number of instruments which had to be administered individually were tried. If they met other criteria then whether they could be either converted into group-administered tests, or whether paraprofessionals could successfully administer them was considered. Three major sources of unreliability were found, including instruments previously reported as reliable: examiner reliability, scoring reliability, and reliability of interpretation of differences. The vocabulary subtest of the Wechslar Intelligence Scale for Children, as improved by Jastak and Jastak, is held to have so far been proven the best single subtest predictor. The visual sequential memory subtest of the Illinois Test of Psycholinguistic Ability (ITPA) was found to predict first grade achievement more powerfully than the total ITPA plus the Stanford-Binet IQ Test. (Author/JM)

PROBLEMS IN THE PRACTICAL ESTABLISHMENT OF PREDICTIVE MEASURES IN THE SCHOOLS: PART TWO

Roger A. Severson
University of Wisconsin

Paper presented at an annual meeting of the American Educational Research Association. This paper formed part of a symposium on Problems in early school detection of learning and problem behaviors, March 4, 1970, Minneapolis, Minnesota.

In another paper Henry Kaplan has documented some of the practical problems encountered in a longitudinal study undertaken with this investigator. These problems were encountered in the attempt to employ teachers as raters of children and as administrators of group tests. This report extends to a discussion of problems encountered in a longitudinal study now in its fourth year where the focus has been the early identification of later learning disorders. The first wave consisted of individual and group testing of 400 first grade children in 1966-67. Subsequently, each year has seen the introduction of new tests, new populations, or new techniques in testing the children. In this report the primary focus will be on practical problems which were found, although some preliminary report of results will also be included. Although many practical problems are invariably found with this kind of study, only those problems which seem to have implications for other settings will be described here. The audience for whom these remarks are intended are persons in other systems who are attempting to set up practical detection programs as a way of improving intervention programs or general school programing in the primary grades.

The general goal which guided this study was the identification of tests with the characteristics of high reliability, low cost, short time to administer, low demand on scoring sophistication, and which could be given in group form by the classroom teacher. However, we tried out a number of instruments which had to be administered individually. If they met other acceptable criteria, such as improving the effectiveness of prediction, then we considered if they could be either converted into group-administered tests, or if paraprofessionals could successfully administer them.

The first major problem encountered was that of reliability. Since

four examiners were involved in the individual administrations in the first year, an analysis was done of inter-examiner differences on the IQ subtests employed. Significant differences were found between cumulative averaged scores, and this finding has led us to investigate more extensively the consistent variations between examiners. Since one examiner was myself and the remaining three were graduate students who had all completed an IQ course under me, the variations are probably not representative of the range which would be encountered among random examiners, regardless of level of training.

In a recent effort to examine this area more fully, we had two examiners test 20 children each who were randomly drawn from two kindergarten classes. One was a quiet male examiner who seldom emitted social praise of an excessive amount. The other was a female examiner who emitted high frequencies of effective social reinforcement. Tabulated subtest profiles and total test scores revealed a striking finding: on all subtests where verbal expression was important for answering, the female examiner invariably obtained higher scores. On the WPPSI, whereas Performance IQ scores did not differ significantly, there was an average of 10 points difference on the Verbal IQ scores! Although we are cautious about generalizing too widely, because of the rather small number of examiners involved, I am convinced we will find rather impressive differences between examiners when we begin to look carefully at this area.

As if these differences in examiner reliability were not upsetting in themselves, we have found two other major sources of unreliability in our work. We recognized that score reliability in the above study could be due either to examiner effects (whether due to cueing, reinforcement, or reactions generated in the child as a function of sex and bearing of the examiner), or to scoring and interpretation of differences. We took a number of Bender Gestalt protocols and had several graduate students score them independently with Koppitz' system (1964). The results, contrary to

a couple of reassuring studies in the literature, resulted in intra-item reliability in the .30s! We found the deceptively stable total scores were too frequently achieved by including a number of items which achieved only random agreement across scorers, making the total scores spuriously reliable. The results of this led to an extensive revision of the Koppitz system in order to achieve higher reliability at the item level, an activity which may be ironic in that we have increasingly questioned the appropriateness of the Bender Gestalt Test as a truly useful diagnostic tool. We have also looked at the Verbal Expression subtest of the ITPA, which seemed to present scoring problems, and found the same shocking unreliability. It should be emphasized that we were using graduate students with limited experience, but who exercised great care in scoring. Frankly, I think they were more careful than the average school psychologist who administers tests as a routine part of his work. The literature has also recently begun to look at scoring reliability on the IQ tests with very unreassuring results.

The problems of unreliability, unfortunately, do not stop with those of examiner administration and scoring. The fact is that the child in the primary grades is a very changeable individual. In fact, whenever I have been able to get true test-retest figures on tests given at this level, although test-retest with one day intervals typically yield the encouraging reliabilities we like to see in the literature (180 or higher), with intervals of 60 days the reliabilities all seem to fall off into the .50s or .60s. Although this isn't too surprising in itself, it should alert persons who match detection programs given at one point with intervention strategies started several weeks later. The effect is particularly pronounced for screening tests which provide a small spread of scores. The answer is fairly simple; lengthen the sample of items which differentiate. This can be done either by adding items to a battery, or by giving the test two different times. An advantage of giving a screening test at two

different times, such as at the end of kindergarten and the beginning of first grade, is that it not only lengthens the number of items given but it allows for plotting children who are undergoing impressive change in a given area of functioning.

In addition to investigating the shocking problems of unreliability in our own tests, we looked at the existing instruments used by teachers for identifying "readiness" for reading and arithmetic. It was rather comforting to find a similar amount of unreliability upon randomly checking several protocols. A summary of the general value of readiness tests can be made by the finding that predictively the subtest which tests for knowledge of letters is almost always as highly related to later success in first grade as any test using more extensive testing. In agreement with the work of Wilma Hirst, (1969), we found the arithmetic subtest of the Metropolitan Readiness Test to be a good predictor, and concluded it was largely due to the broad representation of concepts in this subtest. It is, in effect, an academically disguised IQ test.

Over the past four years we have had a chance to look at several of the popular instruments in a variety of ways. Another shocking finding is the existence of the myth that IQ tests predict learning ability. When we began to find that certain IQ subtests did not relate at all to achievement, we returned to the literature to see what empirical support had been created earlier. We found an incredible lack of predictive validity studies in the literature. In fact, in one impressive review of the area of reading readiness, we found the generalization that ". . . everyone knows intelligence relates about .60 with achievement in the first grade." (Chacko, 196). We established that only in recent years have any investigators (such as Hirst and Dudek) actually begun to publish empirical data on the true predictive validity of the IQ tests in the early grades. Their value, as they have attested, are greatly overplayed.

What is not yet apparent is that the same findings will probably accrue for many of the major tests used by psychologists to predict learning ability. Our experience with the Frostig Developmental Test of Visual Perception is that it is not as powerful as the readiness tests in predicting later learning. The same thing seems to be happening in our studies with many other popular tests, such as the ITPA, the Bender Gestalt Test, the WPPSI, and several lesser known tests. However, the effect is not total. That is, when we look at a battery test like the ITPA we find a single subtest with impressive power, and the same is true for the IQ tests.

If we consider for the moment the value of homogeneous subtests, we have identified some of genuine potential value. For example, the vocabulary subtest of the WISC has so far proven to be our best single subtest predictor, but not until we found someone who had improved it. Jastak and Jastak (1963) were not impressed with the subtest and they revised the number of items (from 40 to 25), reordered them by difficulty level, and published a marvelously improved scoring manual. Another "find" was the visual sequential memory subtest of the ITPA. Hirshoren (1969) found this single subtest to predict first grade achievement more powerfully than the total ITPA plus the Stanford-Binet! As it happens, we created our own visuoperceptual memory task before learning the value of this ITPA subtest. It also proved to be a powerful predictor in our work. Since the vocabulary subtest and these visuoperceptual memory tasks have proved in two studies to be virtually unrelated, and to both be powerful predictors, we are on our way to maximizing the purely predictive variance which can be obtained. Although not yet replicated, and certainly not yet compared with a variety of other pretenders to the throne, these two variables relate in the .80s to achievement one year later. In the face of all that unreliability mentioned earlier (which these both partly overcome), this is impressive indeed.

Several other instruments with either predictive or diagnostic value

should receive brief comment. We view the parents as important sources of information, and sources which can aid intervention programs prior to school contact. Unfortunately, our data on parent-completed inventories prior to school experience have not yet been analyzed, but when we created an inventory and administered it to parents in late first grade, the relationships with school achievement were in the .70s, and this was when we had partialled out between-classroom differences and individual IQ scores. This is pretty impressive as a beginning. We did learn that the inventory we used generated more parent negativism than we wished. We initially avoided the purely socioeconomic information (education, income, house value) on the grounds this would be resisted more, but now we feel it would be wiser to ask about education levels and occupation, and a few questions about interests of the child, habitual ways of reacting, and other neutral questions. Despite the value of parental response, the acceptance of the testing procedure must be carefully checked before drawing on their knowledge. We have learned we can reassure parents by telling them exactly how we intend to use the information, but we view this part of the early detection movement as quite controversial.

Two areas where we are still actively seeking better instruments is in the measurement of auditory and visual discrimination. The two pre-eminent tests, the Wepman Test of Auditory Discrimination, and the Frostig DTVP, have been grossly disappointing. Our average Wepman scores have exceeded the norms of clinical concern as reported by Wepman, and we have no real belief that the test relates to actual language comprehension as it occurs in the classroom. The area needs much more investigation in a criterion-oriented way. We are currently impressed with the Alameda County, California, work in this area, (Lasnik, 1970) since they also have remedial tapes available, and hope to by-pass the Wepman test completely. In the area of visual perception, what we hope to find is a test comparable to the Beery-Buktenica

sequence on visuomotor abilities, known as the Visual Motor Integration (VMI) test. The logic which underlies this particular test, well articulated by Keith Beery in his monograph written for Follett Corporation (1967), has guided us in most of our approaches to the problems of detection-intervention.

In the effort to systematically work through the available measures used in early prediction, both from the areas of education and psychology, it did not dawn on us until fairly recently that most tests confound two basically dissimilar behaviors, just as reflected in the aptitude and academic content mixture found in the Metropolitan Readiness Test. We now distinguish between knowledge of specific curriculum content and the various discriminations and acquisitions necessary to master new learning. The IQ test is a good example of a test which confounds old learning and current learning abilities and gives no really useful way of separating them. We have found it much more useful to give a curriculum pretest, such as knowledge of letters, words and number operations directly related to the specific curriculum being used. Then we also check out the ability to the child to master new learning. Since our inspection of current teaching practices and curriculum sequences in the primary grades has convinced us that serial and paired associate learning is by far the most common kinds of learning, we use these kinds of tasks to assess the child's capacity to learn new material. Using a learning and memory model, with auditory and visual acuity and discriminaiton as important input behaviors, we feel we are on the right track for establishing the kinds of powerful detection techniques leading to truly effective intervention programs. By working backwards from this criterion-oriented model to procedures which are as inexpensive as possible, which can be given in group form by the teacher or individually by a paraprofessional, which provides acceptably reliable estimates of the behavior, and which lead to behaviors which can be changed effectively, economically and with proven benefit, we feel we are developing practical procedures with a high probability of community acceptance and adoption.

Although the major focus of our work has been on examining the value of test instruments completed by the child, we fully recognize that these behaviors are mediated through a classroom interactive process. We have sought to extend our investigation of these instruments to their selective value in individual classrooms. As yet we have not been able to look at the cumulative interactions between incoming child with particular behaviors, teacher with her own repertoire of behavior, and the reactions of child and teacher to each other. We have, however, looked at the differences of groups of children from classroom to classroom. What we are impressed with is the variability of ability measures from one classroom to another. National norms are worthless. When we examined 19 different classrooms on 12 different variables we found marked differences. We have rediscovered individual differences on a classroom level. Although we aren't that far yet, it may very well mean that predictors change from classroom to classroom in their importance. In a given classroom in a given year it may be more important that you can speak glibly than that you can differentiate figures on a page. Or a teacher may reinforce short-term rote retention as opposed to creative responding. The complex interactions certainly lie before us, but we should not despair simply because of the complexity.