DOCUMENT RESUME

ED 128 415 TM 005 599

AUTHOR Epstein, Kenneth I.
TITLE An Empirical Investigation of Four Criterion-Referenced Testing Models.
PUB DATE [Sep 75]
NOTE 14p.; Paper Presented at the Annual Conference of the Military Testing Association (17th, Fort Benjamin Harrison, Indiana, September 15-19, 1975); Also included in TM 005 585

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Bayesian Statistics; *Criterion Referenced Tests; *Mathematical Models; Military Personnel; Probability; *Raw Scores; *Statistical Analysis; Test Interpretation; *True Scores
IDENTIFIERS Army; Rasch Model

ABSTRACT

Since the primary purpose of classical testing is to rank order examinees consistently, the absolute value of the true score has been relatively unimportant. However, the major purpose of criterion referenced testing is to estimate the true capabilities of examinees to perform specific tasks. Hence, the problems of true score determination assume critical importance. Four measurement models which have potential for evaluating the results of criterion referenced tests are discussed here. The Proportion correct model assumes that the proportion of sample trials scored correct is an unbiased estimate of the proportion correct in the infinite domain for that individual. The binomial error model adds the specification of the conditional distribution for the observed score for the given true proportion correct. The third model applies the philosophy implied in the binomial error model to Bayesian statistical theory, and the final model is the Rasch one parameter logistic model. Advantages and disadvantages of each are discussed, but the final choice of a model is to be based on the needs of the testing program and the resources available to analyze the data. (BW)

An Empirical Investigation of Four Criterion-Referenced Testing Models

Kenneth I. Epstein
Army Research Institute for the
Behavioral and Social Sciences

For the past two years, the Army Research Institute has been engaged in basic testing research under Project METTEST - Methodological Issues in the Construction of Criterion-Referenced Tests. Project METTEST was conceived to provide basic support for the Army's rapidly growing trend towards performance oriented training and testing. This paper summarizes part of the thinking which has evolved from METTEST.

One question that is present whenever test scores are considered is, What are the "true" scores that the fallible test scores represent? The problem is so complex that there is considerable discussion about what "true" score means (see, for example, Lord and Novick, 1968 pp. 39-44). Even if a definition is agreed upon, the estimated value of the true score given an observed test score will vary according to the particular measurement model used to evaluate the data.

The absolute value of true score has been relatively unimportant for classical measurement. This is because the primary purpose of classical testing is to rank order examinees consistently. For this purpose, the critical problem is not determining the true score, but rather it is determining the correlation between true scores and observed scores. The techniques of classical measurement are powerful because it is possible to determine this correlation without knowing the actual values of the true scores. It is possible to evaluate how well a given test yields scores that correlate highly with true scores and hence, how well the test will rank order examinees consistently.

Criterion-referenced testing does not allow for the luxury of placing the emphasis in test evaluation on correlation. In fact, it is possible for a good criterion-referenced test to correlate zero with true score. This might occur in a mastery learning context where all examinees have attained an equal level of ability and all examinees obtain the same test score. If one attempts to determine the correlation between a

The views expressed in this paper are those of the author and do not necessarily imply endorsement by the U. S. Army.

collection of equal observed scores and equal abilities, the value of the coefficient will be zero. The major purpose of criterion-referenced testing is not to rank order examinees consistently. Rather, it is to estimate the true capabilities of examinees to perform specific tasks. Hence, the problem of true score determination assumes critical importance.

This paper discusses four measurement models which have potential for evaluating the results of criterion-referenced tests. In particular, the paper compares the estimates each model yields for true scores given observed scores. The four models to be discussed are 1) a model based on observed proportion correct, 2) a binomial error model, 3) a Bayesian model, and 4) the Rasch one parameter logistic model.

The data for the analyses were collected as part of a study which evaluated tank gunnery training devices (Rose, et. al, 1975). The data consist of hit/miss scores recorded for 12 rounds fired from the main gun of the M60A1 tank at a moving target. 154 crews participated in the experiment. A summary of the data is given in Table 1. For the original experiment the 154 crews were broken into seven groups corresponding to different training programs. The reanalysis of the data for this paper ignored the difference in training since it is irrelevant to the present problem. At some future time it might be of interest to study whether test results are consistent across apriori defined different examinee populations.

The major requirement of a criterion-referenced test is that all items come from a well-defined domain (Millman, 1973). In other words, it is essential that all items in the test (or sub-test) measure the ability to perform the same objective. The twelve essentially identical rounds provide an ideal representation of a criterion-referenced test designed to evaluate gunner ability to hit moving 16' x 10' plywood tank silhouettes at ranges of approximately 700 and 750 meters with standard 105mm HEAT TPT ammunition. The twelve rounds represent a sample of an infinite number of similar rounds that could have been fired. Thus, it seems reasonable, and it is consistent with established measurement theory, to define the true score as the proportion of hits that would be obtained were the infinity of rounds fired. In other words, the twelve rounds represent a sample of trials chosen from an infinite population of possible trials. The task then is to infer from the performance on the twelve trial test what proportion of hits would be achieved if all trials were given. The following section of the paper discusses four procedures for accomplishing this task.

The first procedure uses the obtained proportion correct as the estimated true score. If we assume that the test is a random sample of trials from the domain of interest, and if the trials are dichotomously scored, and if we assume that the response to any given item is not dependent on the response to any other item (e.g., there is no learning occurring during testing, nor is there any fatigue or similar effect occurring), then it is appropriate to describe the data mathematically as a series of independent Bernoulli trials. In this case, the proportion of sample trials scored correct by an individual is an unbiased estimate of the proportion correct in the infinite domain for that individual. This model implies that all items are of equal difficulty for an individual. For example, if the proportion that would be correct in the whole domain is p for some individual, then the probability that he would respond to any randomly sampled trial correctly is also p. Notice that this does not imply that two individuals of differing ability will find the trials equally difficult. The more capable individual will find trials uniformly easy, the less capable individual will find the same trials uniformly difficult. Table 2, Column A shows the proportion correct corresponding to each obtained score on this 12 trial test.

The proportion correct can also be shown to be the maximum likelihood estimator of the true proportion correct (Lord and Novick, 1968, p. 88). The proportion correct has a mean value (over repeated sampling) of p, the true proportion correct and a variance of p(1-p)/n. Hence, particularly for small sample sizes, this estimate of the true score is probably not adequately reliable. Notice that the variance of the estimated proportion correct will not be a constant for all obtained scores. In fact, it is largest for the mid range of the distribution (p = .500 yields the largest variance, .25/n), and decreases to zero at the extremes. This is not an unreasonable result. We would expect very good or very poor examinees to respond in a highly predictable and consistent manner. It is far more difficult to make fine discriminations near the middle of the distribution. Hence, the practice in test construction to design items to be most sensitive to the range of abilities near the middle. Lord and Novick (1968) point out an interesting paradox associated with this type of analysis.

> "The standard error of measurement is smallest for examinees whose true scores are nearest one or zero. Should it not follow from this that the best measuring instrument is a test composed of items so easy that everyone will have a relative true score near one, or so hard that everyone will have a relative true score near zero?
>
> The answer to this question is that the effectiveness of a test as a measuring instrument usually does not depend merely on the standard error of measurement, but rather on the ratio of the standard error of measurement to the standard deviation of observed scores in the group. The more discriminating the test items, the larger will be the standard deviation of observed scores, other things being equal; and hence, the less will be the

> danger that true differences will be swamped by random errors of measurement and lost to view.
>
> The small standard errors of measurement that result when a test is made very easy are not beneficial because the standard deviation of observed scores for such tests is also small. This is most apparent in the limiting case when the test is so easy (or so difficult) that everyone gets a perfect (or a zero) score and both standard deviations are zero. Even though in this case there are no errors of measurement at all, such a test obviously is not discriminating among examinees and thus is not a useful measuring instrument (p. 252)."

One possible exception to the Lord and Novick statement might occur if a sample of examinees all of whom were complete masters (nonmasters) of the material took the test. Then one would have a test that did not distinguish among identical individuals, which is what is desired. However, the occasions when the examinee population is likely to be homogeneous enough for this to occur are probably infrequent.

A natural extension of the proportion correct model is the binomial error model. The binomial error model is more powerful than the simple proportion correct because the entire distribution of observed responses is included in the analysis. All of the assumptions discussed in relationship to the proportion correct model hold for the binomial error model. The major addition is the specification of the conditional distribution for observed score x for given true proportion correct T. This distribution is the binomial:

$$(1) \quad h\,(x|T) = \binom{n}{x} T^x (1 - T)^{n-x} \quad x = 0,1,\ldots n., \quad 0 \leq T \leq 1,$$

and n equals the total number of trials on the test.

Lord and Novick (1968) prove a very useful consequence of the model. "Under the binomial error model, if the observed score distribution is negative hypergeometric, then the regression of true score on observed score is linear (p. 517)." They then outline a procedure for estimating the parameters of the negative hypergeometric distribution from observed scores. The procedure was carried out for the tank gunnery data and the theoretical distribution was compared to the observed distribution. The value of X^2 for this analysis was 3.451. Evaluation of this value with nine degrees of freedom yielded $.95 > p\ (X^2 = 3.451) > .90$. Since this represents adequate fit it is possible to proceed with the analysis assuming that the regression of true score on observed score is linear.

The regression function can be written

(2) $E(T|x) = \alpha_{21}\, x/n + (1 - \alpha_{21})\, \mu_x/n$, $x=0,1...n$, and μ_x = the mean of the observed scores, $\alpha_{21} = n/(n-1)\, [1 - \mu_x(n-\mu_x)/n\sigma_x^2]$ σ_x^2 = the variance of the observed scores. (Lord and Novick, 1968 p. 517, 521)

For these data the value of the regression function is

(3) $E(T|x) = .059\, x + .161$, $x = 0, 1, ...\, n$.

The estimated true score calculated using the above regression function are found in Table 2, Column B.

Comparing these results with those obtained under the proportion correct model shows that they are comparable, particularly in the mid range of the distribution. The differences are directly attributable to a regression effect in which the extreme values are regressed toward the mean of the distribution. What is happening, in effect, is that performance of the group is being used to help temper judgments about individuals. On the one hand, this point of view seems to contradict the philosophy underlying criterion-referenced measurement, that an individual should be judged on the basis of his ability and not compared with his peers. However, since in most cases the examinee population does have some characteristics in common and since extreme scores are suspect it makes sense to use all available data. This particular model is especially attractice because it has a built in validity check. If one is not successful in fitting the negative hypergeometric distribution to the data it implies that the regression is either not linear or that the binomial error model is not appropriate. In any case, it immediately warns the user to be careful of any interpretations he makes.

Lewis, Wang, and Novick (1973) have applied the same philosophy implied in the binomial error model, that all available information be utilized, to the development of a measurement model based on Bayesian statistical theory. The procedures are complex and require computerization for full utilization. The results of using their procedures to evaluate these data reveal some interesting and thought provoking implications of the Bayesian approach.

The procedure begins by mapping the observed scores into a new set of variables (g_j) using an arcsine transformation. The g_j are assumed to be normally distributed with mean $\gamma_j = \sin^{-1} \sqrt{T_j}$ and variance $\gamma_j = (4n + 2)^{-1}$, where γ_j is the transformed value of the true proportion of successes, T_j, and n in the number of test items. The assumption of normality is shown to be reasonable for tests of at least eight items. In addition to the observed data, the procedure requires that two additional parameters be specified by the user. These parameters describe the prior beliefs concerning the distribution of γ_j. The γ_j are assumed to be a random sample from a normal distribution with mean μ_Γ, variance ϕ_Γ The μ_Γ and ϕ_Γ are assumed to be independent, having a uniform and inverse chi-square distribution respectively. The first additional parameter that must be specified is the degrees of freedom for the inverse chi-square distribution. A recommended value for most practical purposes is eight. This value was used for the analysis described in this paper. The second additional parameter is also related to the inverse chi-square distribution. It is designated t and can be thought of as the length of a test that the user would consider to offer as much information as he now has (before testing) about the examinees. Thus, if very little is known, t will be small, and if the prior information is extensive, t will be large. Since relatively little prior information was available for the data described in this paper, the value of t chosen was three. Clearly, before this procedure can come into wide use, the relative importance of the t value to the final results must be investigated. Experience based on empirical applications of the procedure will also help in establishing practical guidelines for the use of the procedure. (For a more detailed discussion of the rationale behind this procedure see Novick, Lewis, and Jackson, 1973 and Lewis, Wang, and Novick, 1973).

Once the observed data have been transformed and the parameter values specified, the application of the model is relatively straightforward. Two alternative procedures are available. For test lengths up to 30 items, examinee groups up to 80 persons, and transformed score variances up to .05 (reasonable values for most applications), Wang (1973) has prepared tables of constants for carrying out the necessary calculations for their use. For larger sample sizes, Lewis, Wang, and Novick (1973) provide a procedure for carrying out the necessary calculations which involves solving a cubic equation. The later procedure was used for this example.

The most important result of the procedure is a regression equation for the posterior values of γ_j. For the example data the equation is,

(4) $E(\gamma_j \mid \phi_\Gamma, g) = .710\, g_j + .244$, where ϕ_Γ is the posterior variance

solved for by a cubic equation and the g_j are the transformed variables.

It is interesting to compare this equation with the regression equation obtained as a result of applying the binomial error model (Equation 3). The two equations have essentially the same form. While they cannot be directly compared since the Bayesian equation is written in terms of transformed variables one can compare the relative weights given the mean of observed data and the individual variables themselves. (The pure number in each equation reflects the weight applied to the mean). In the Bayesian approach the individual variables are weighted much more heavily than the mean. In the binomial error approach the mean seems to be relatively more important. This should result in the true proportion estimates found by the Bayesian approach to be regressed less toward the mean than the binomial error model results. These results are shown in Table 2, Column C. In general, a regression effect is seen for the Bayesian results but it is not as strong as the binomial error model results.

The heavy weight accorded the individual score values in the Bayesian approach is a direct result of the small amount of prior information. This makes intuitive sense, for if little is known about a group it seems unreasonable to put much emphasis in overall indices of group ability such as the mean. Had the prior information been more conclusive, more use would have been made of group data in determining the estimates of the true porportions correct.

The final model to be discussed is the Rasch one parameter logistic model. Superficially the Rasch model appears to be conceptually very different from the previously discussed models, however the differences are more apparent than real. The Rasch model hypothesizes that people are distributed on an underlying ability trait and further that their response to a test trial or item is governed purely by the ability of an individual and the difficulty of the trial. This is analogous to the probability of responding correctly to any given trial that underlies the response patterns of examinees in the proportion correct model, the true score distribution that is observed as a negative hypergeometric distribution in the binomial error model, and the posterior distribution of abilities in the Bayesian model. The Rasch model's strength lies in the fact that it is possible to calibrate a set of items with differing difficulties and administer different subsets of those items to different groups of examinees. The resulting estimates of individual abilities will be on the same scale and it will be possible to compare individuals regardless of the particular items chosen.

The data in this example were analyzed at the University of Chicago by Dr. Benjamin Wright. The results of the analysis showed that the twelve trials differed in apparent difficulty. The first trials seemed to be more difficult than the latter trials indicating that a gradual learning effect occurred during the test. Notice that this implies that the assumption that all items were identical is <u>not</u> strictly valid. It also calls into question the validity of interpreting the true score as the proportion of items that would be correct if all items were given. If items are not equally difficult is it not more important to know which items are responded to correctly? On the other hand, if this level of specificity is desired it will require far more extensive testing than seems practical. Two alternatives seem available. The first is to accept the approximation implied by the definition of true score as the proportion correct for all items. For most practical purposes this seems to be acceptable. The other alternative is to begin thinking in terms of latent traits. Such a transition will require that a large amount of empirical data be collected so that abilities expressed in terms of latent traits cna be given meaning in terms of observable behavior. For example, the data discussed in this paper were calibrated according to the Rasch model so that individuals scoring six hits were assigned an ability value equal to 0.00, and those scoring eleven hits an ability value equal to 2.45. (The entire set of Rasch ability values is shown in Table 2, Column E.) Stating abilities and minimum standards in terms of latent variables will allow for great flexibility in testing, but will require major efforts in interpretation.

The major purpose of this paper is to compare estimated true scores obtained by applying several measurement models. Therefore, it was necessary to sacrifice some of the information obtained from the Rasch analysis in order to obtain values on the same scale as the other values. This was accomplished by assigning the average difficulty of the twelve trials to each trial and applying the basic equation of the Rasch model.

The Rasch model states that the probability of a correct response to any given trial by a given individual is a function of the difficulty of that trial and the ability of the individual:

(5) $p = e^{(b-d)} / 1 + e^{(b-d)}$, where b is the item's difficulty and d is the person's ability. If the above equation is solved with b equal to the average item difficulty, the estimated average probability of a correct response is found. This value corresponds to the estimated true proportion correct found using the other models. These results are shown in Table 2, Column D. (Note that results are not shown for zero correct or twelve correct. This is because neither extreme score is used in the calibration and hence no ability estimates are obtained.) The estimates for the Rasch model are very similar to those found for the proportion correct model. This occurs because the Rasch model calibration attempts to find values for b and d which duplicate the observed data. The reliance on the group mean which is incorporated in the binomial error model and the Bayesian model is not found in the Rasch model.

Comparison of the estimated true proportion of hits for the four measurement models indicates relatively little difference among the models. For practical purposes, such as assigning individuals to mastery groups, there are not likely to be great differences when different approaches are used. Until more theoretical and empirical work has been completed it is not possible to make qualitative statements about the different approaches. However, some general strengths and weaknesses of the models can be identified.

The proportion correct model is clearly the easiest to apply. Calculations are minimal and the interpretation is straight forward. The approach has two weaknesses. First, none of the information about the group is incorporated. Group data is valuable in interpreting test results and should be considered. A second related weakness is that there is no direct connection between the estimated true proportions correct and the observed data. The calculations for the proportion correct model can be done without observed data. Thus, it seems to offer a good first approximation. If the observed data make sense when interpreted in terms of this model then it can probably be utilized. However, if examinees perform very differently than they would be expected to perform, then the model may be inappropriate.

The binomial error model shows, in effect, what happens when the proportion correct model is applied to observed data. Its strengths lie in its use of all the test information and the built in check on its fit to the data. An obvious weakness is that the model cannot be used if the negative hypergeometric distribution does not fit the observed data. In such cases the regression of true score on observed score is not linear. Techniques do exist for calculating the nonlinear regression but they require smoothing, estimation procedures, and tedious computations. Whether such measures are warranted is questionable.

The Bayesian approach will be most useful when significant prior information is available. For cases such as that presented in this paper the true power of the Bayesian approach will not be demonstrated. The Bayesian approach has the advantage of incorporating prior information into the analysis in addition to incorporating all the observed data. The major disadvantage of the Bayesian approach is that very often intuitive estimates of priors are in error. Such errors can lead to misleading results. However, the potential of the Bayesian approach for increasing the precision and efficiency of testing warrants that practical guidelines for its application be developed.

The Rasch model presents an opportunity for testing to pursue new directions. It has the potential for greatly increasing flexibility in testing. Like the binomial error model and the Bayesian model, the Rasch model relies on observed data in calculations, however once a calibrated item pool is available the person ability estimates are free of the particular item set used. This independence from the item set puts the major emphasis on the individual's ability. Thus, it seems philosophically more attuned to criterion-referenced testing. Weaknesses of the

Rasch model include possible problems in interpretation and the fact that, not all data sets will fit the model. If the data do not fit, it requires major revisions of the test or recourse to an alternative model.

The final choice of a model should be based on the needs of the testing program and the resources available to analyze the data. It is hoped that this paper will lead practitioners to more carefully consider their test results, whatever model they choose. In this way, perhaps the interpretations of test scores and the decisions based on them will be improved.

Table 1
Summary of Tank Gunnery Data

Observed Number of Hits	Frequency	Round Number	Proportion of Examinees Scoring a Hit
0	2	1	.429
1	2	2	.487
2	10	3	.526
3	12	4	.474
4	13	5	.500
5	16	6	.558
6	20	7	.545
7	16	8	.662
8	19	9	.578
9	14	10	.636
10	14	11	.604
11	11	12	.630
12	5		
	m=154		

Mean Number of Hits: 6.630
Variance: 8.457
Mean Proportion of Hits: .553
$KR\text{-}20 = k/k-1(1- \Sigma pq/\sigma_T^2) = .7136$

Note: The KR-20 reliability coefficient is included here not because the author necessarily advocates its use in evaluating criterion-referenced tests. It is included because it helps to describe the nature of this data and this group of examinees. A paper discussing some valid interpretations of classical measurement techniques for criterion-referenced tests is in preparation.

Table 2
Summary of Estimate True Proportion of Hits

Observed Score	A: Proportion Correct	B: Binomial Error	C: Bayesian Model	D: Rasch (Proportion)	E: Rasch (Ability)
0	.000	.161	.093	----	----
1	.083	.220	.201	.079	-2.45
2	.167	.279	.278	.161	-1.65
3	.250	.338	.342	.244	-1.13
4	.333	.397	.401	.330	-0.71
5	.417	.456	.469	.416	-0.34
6	.500	.515	.516	.500	0.00
7	.583	.574	.574	.584	0.34
8	.667	.633	.633	.670	0.71
9	.750	.692	.691	.756	1.13
10	.833	.751	.752	.839	1.65
11	.917	.810	.821	.921	2.45
12	1.000	.869	.921	----	----

References

Lewis, C., Wang, M.M., & Novick, M.R. Marginal distributions for the estimation of proportions in m groups. ACT Technical Bulletin No. 13. Iowa City, Iowa: The American College Testing Program, 1973.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company, 1968.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Novick, M.R., Lewis. C., & Jackson, P.H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.

Rose, A.M., Wheaton, G.R., Leonard, R.L., Fingerman, P.W. & Boycan, G.G. Evaluation of two tank gunnery trainers. Arlington, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences, in press.

Wang, M.M. Tables of constants for the posterior marginal estimates of proportions in m groups. ACT Technical Bulletin No. 14. Iowa City; Iowa: The American College Testing Program. 1973.

Wright, B. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.