DOCUMENT RESUME

ED 113 376 TM 004 873

AUTHOR Moonan, William J.
TITLE Contributions to Objective Measurement and Evaluation of Trainee Competency.
PUB DATE Oct 74
NOTE 33p.; Paper presented at the Conference of the Military Testing Association (Oklahoma City, Oklahoma, Oct., 1974)

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS Computer Programs; Goodness of Fit; Item Analysis; *Mathematical Models; Matrices; *Measurement Techniques; *Military Personnel; *Military Training; *Personnel Evaluation; Probability; Statistical Analysis; Systems Development; Trainees
IDENTIFIERS Binary Measurement Model; *COMET; Frequency Ratio Method

ABSTRACT

The purpose of this paper is to lay a basis for and discuss the components of a system, called COMET, designed to objectively measure and evaluate the competency of trainees in military training enterprises. COMET is an acronym for "Computerized Objective Measurement and Evaluation of Trainees." These goals will be accomplished by: (a) describing what the author means by the terms "measurement," "objectivity," and "evaluation," by indicating systems antecedent to COMET, and describing the objective Binary Measurement Model (BMM) used in the COMET system, (b) deriving and illustrating a new procedure for estimating one of the parameters and fit of the BMM model, called the Frequency Ratio Method (FRM), and (c) specifying the component subsystems, and their functions, of COMET as well as its goal, objectives and practical problems associated with its implementation. A detailed numerical example is provided to facilitate understanding of COMET's mathematical and statistical aspects. (Author)

ED113376

CONTRIBUTIONS TO OBJECTIVE MEASUREMENT AND EVALUATION OF TRAINEE COMPETENCY

by

William J. Moonan, Ph.D.

Director, Statistical Programs
Navy Personnel Research and Development Center
San Diego, California 92152

A Presentation Made to

The 1974 Military Testing Association Conference
Oklahoma City, Oklahoma

21-25 October 1974

393

TM004 873

TABLE OF CONTENTS

LIST OF EXHIBITS

1. Introduction

The purpose of this paper is to lay a basis for and discuss the components of a system, called COMET, designed to objectively measure and evaluate the competency of trainees in military training enterprises. COMET is an acronym for "Computerized Objective Measurement and Evaluation of Trainees."

These goals will be accomplished by:

a) describing what the author means by the terms "measurement," "objectivity," and "evaluation," by indicating systems antecedent to COMET, and describing the objective Binary Measurement Model (BMM) used in the COMET system,

b) deriving and illustrating a new procedure for estimating one of the parameters and fit of the BMM model, called the Frequency Ratio Method (FRM), and

c) specifying the component subsystems, and their functions, of COMET as well as its goal, objectives and practical problems associated with its implementation.

A detailed numerical example is provided to facilitate understanding of COMET's mathematical and statistical aspects.

2. Background

By the term "measurement" the author refers to the process by which numbers are assigned to a property of an entity. In training enterprises, the entities are personnel or trainees whereas the property being measured we refer to as "competency" in the training field.

By the term "evaluation" the author refers to the process by which measurements or their functions are used to characterize features or aspects of the evaluated "situation" or the entities included therein.

For example, we may evaluate a trainee (entity) during training (situation) by the process of taking the difference (function) of his competency (measurement) as determined at two different stages of training. Notice that no other value judgment is placed on the evaluation result in this definition. This was omitted because the value applied to an evaluation result is not unique. It is different depending upon the decision maker and the decision problem wherein the evaluations are to be utilized for decision purposes. If the results are not to be utilized for decision purposes, they need no value judgment component.

By the term "objective" the author means different things depending upon what is referred to. There are three different senses in which the term is used in this paper. They are

a) Training program objectivity,

b) Measurement Instrument objectivity, and

c) Measurement model parameter objectivity.

Training Program Objectivity refers to the process whereby training programs are objectively established by means of expressly stated goals, functions and objectives and are designed and managed by a carefully thought out, systems approach such as that developed by Rundquist (1967).

Measurement Instrument Objectivity refers to process whereby a measurement instrument (test) is designed to be free of ambiguity and subjectivity in its administration or in responding to or scoring its items.

Measurement Model Parameter Objectivity refers to two properties of a measurement model, viz.

a) the comparison of any two testees may be carried out in such a way that no other measurement parameters are involved than those of the two testees, and

b) any two items may be compared independently of all others parameters (including those of the testees) than those of the two items.

We emphasize the desirability of requiring an evaluation system, automated or not, to possess the three aspects of objectivity, otherwise it is likely not to meet its goal and objectives as well as not providing useful measurements and evaluations.

As far as is known, there is only one class of measurement model (those developed by Rasch (1960) which possess the properties of measurement model objectivity. That these properties actually exist for some of the Rasch models, has been shown by Rasch himself (1969), Wright (1969), Choppin (1968), and Schmidt (1969). Later in this paper we will provide yet another demonstration. A most startling and readable account of the Rasch BMM's objective features has been provided by Wright (1968). My own account, Moonan (1972), may be of use and conveniently available to you.

We are aware of at least two antecedent, automated trainee evaluation systems. There are probably others, but our attention has not been directed to them. The designers deserve our applause for their pioneering work and we can only wish that the COMET system we propose could be implemented as effectively and successfully as each of these were. The first system, I refer to as "GLAKES," was developed by CDR W. H. Wheeler while stationed at the Navy Electronic Technician's School at the Naval Training Center, Great Lakes, Illinois. My documentation, Wheeler (1966) is meager, but I did see it in operation. I was particularly impressed by its test reproduction subsystem wherein examinations were quickly and conveniently assembled from the item files. Its evaluation subsystem was also excellently designed.

The second system was developed for the Marine Corps Communications Electronics School at the Marine Corps Recruit Depot, San Diego by Dr. Richard S. Hatch, (Decisions Systems Associates, Inc., Rockville, MD). This system is known by the acronym ATAC; Automatic Testing and Attrition Control and, as its name implies, contained a subsystem for controlling how many, which, and when trainees should be attrited from the training program. This system had other excellent features as well. Most particular were its reporting subsystem and the evaluation subsystem.

Both systems were ahead of their time, the BMM was not well known nor computationally effective at the time of their design. Their development shows that only "bare bones" design features can, at first, be implemented but must ultimately be augmented by features which fulfill local requirements, overcome constraints and meet local desires. It is believed that the basic COMET system, described herein, supplies the essential "bare bones" for many implementation situations.

It should be mentioned here, most fittingly, that my own particular interest in this research area was substantially motivated by seeing GLAKES in operation, working with Dr. Hatch on ATAC and being encouraged by CAPT O. L. Dawson, USCG, to develop an effective COMET-like system for the U.S. Coast Guard Training Facility in New York.

My personal documentation on ATA is limited, Blakely (1969), and Hatch (1967) but, perhaps Dr. Hatch could supply additional information if it is needed.

3. The Binary Measurement Model

Rasch (1960) developed, essentially, three sub-classes of a general class of objective measurement models. I refer to these models as CD, AD and SP models as described in Exhibit 1. Actually the AD model is sufficiently general to encompass the CD model which in turn can be shown to mathematically encompass SP models. The BMM is the AD model and is the one discussed previously (Moonan, 1972). The mathematical statement of the model is expressed in probability terms. The probability of a response, $X(i,j)$ being correct, $X_{(i,j)} = 1$, when made to the ith item, with easiness, $\varepsilon(i)$, by the nth trainee who has competency $\gamma(n)$ is given by the expression, where j=n,

$$(1) \quad P[X_{(i,j)} = 1 \mid \varepsilon(i), \gamma(j)] = \frac{\varepsilon(i) \cdot \gamma(n)}{1+\varepsilon(i) \cdot \gamma(n)} = P_1 .$$

Similarly, the probability of an incorrect response is given by

$$(2) \quad P[X_{(i,j)} = 0 \mid \varepsilon(i), \gamma(n)] = \frac{1}{1+\varepsilon(i) \cdot \gamma(n)} = P_0 = 1 - P_1 .$$

The model is called "Binary" because responses to items are only allowed to be binary or dichotomous. Responses to Affective Domain items are usually n-ary so that the Rasch polychotomous model is more appropriate for measurement under such conditions. In this case, when $n=2$, the CD:BMM model is equivalent to the AD sub-model which we will not discuss further at this time.

The model produces "measurements" because its use permits the assignment of numbers to the "competency" of trainees. It also "measures" the items since the easiness, $\varepsilon(i)$, of each item may be numerically estimated.

The scale of competency has a zero point and an estimable unit of measurement. Thus, the model, with its objective properties provides a suitable measurement basis for the COMET system.

The BMM has two parameters, $\gamma(j)$ and $\varepsilon(i)$. The first we call "competency" rather than the more usual "ability." We prefer the former term for training situations and have avoided the confusion sometimes resulting from the word "ability." The easiness parameter estimated indicates how "easy" or "difficult" an item's correct response is to discern for a testee.

4. Model Objectivity

As was pointed out in earlier sections, Objectivity plays an important role in the COMET system. In this section we shall examine the model objectivity of the BMM by considering the dualistic concepts and implications of measurement model objectivity. These duals are listed below:

Concept No. 1. The item parameters, $\varepsilon(i)$, can be estimated independently and without knowledge of the competency parameters, $\gamma(n)$, of the persons responding to the items.

Implication No. 1. The easiness scale can be calibrated and an easiness can be estimated independently of the persons, and their competencies, responding to the items. This implication is very important for the COMET system for it means we can collect item responses data from whomever it is convenient to do so. Items can be administered in small sets of items to persons of quite divergent competencies. Groups of similar persons are not required to respond to the same items.

Concept No. 2. The competency parameters, $\gamma(n)$, can be estimated independently and without knowledge of the item parameters, $\varepsilon(i)$.

Implication No. 2. The competency scale can be calibrated and a competency can be measured independently of the items used for measurement. Simply stated this means that, if desired, individualized tests for each trainee could be used, but in doing so, each trainee is measured on the same competency scale.

We shall mathematically prove the two concept statements for the BMM below:

According to the model we have the single item probabilities

(3) $$P[X(i,n)=1] = \frac{\varepsilon(i) \cdot \gamma(n)}{1 + \varepsilon(i) \cdot \gamma(n)}$$, and

(4) $$P[X(i,n)=0] = \frac{1}{1 + \varepsilon(i) \cdot \gamma(n)}$$.

Consider next the probabilities associated with the four possible events which can occur if the same person, n, responds to two items, i and j. To this end let,

(5) $D(i) = 1 + \varepsilon(i) \gamma(n)$ and $D(j) = 1 + \varepsilon(j) \cdot \gamma(n)$ be the denominator terms of our probability expressions, then (assuming response independence)

(6) $$P[X(i,n) = 0, X(j,n) = 0] = \frac{1}{D(i) \cdot D(j)}$$,

(7) $$P[X(i,n) = 1, X(j,n) = 0] = \frac{\varepsilon(i) \cdot \gamma(n)}{D(i) \cdot D(j)}$$,

(8) $$P[X(i,n) = 0, X(j,n) = 1] = \frac{\varepsilon(j) \cdot \gamma(n)}{D(i) \cdot D(j)}$$, and lastly

(9) $P[X(i,n) = 1, X(j,n) = 1] = \frac{\varepsilon(i) \cdot \gamma(n) \cdot \varepsilon(j) \cdot \gamma(n)}{D(i) \cdot D(j)}$

The conditional probability that the nth person gets the ith item correct, given that he scored a total of 1 on both item i and j is given by the equation ratio (11) = (8)/(10) where (10) is given by the equation sum (10) = (7) + (8). Thus

(10) $P[X(i,n) + X(j,n) = 1] = \frac{\{\varepsilon(i) + \varepsilon(j)\} \cdot \gamma(n)}{D(i) \cdot D(j)}$, and

(11) $\frac{(8)}{(10)} = \frac{\varepsilon(i)}{\varepsilon(i) + \varepsilon(j)}$ in which we note the $\gamma(n)$ cancels,

and which indicates the truth of Concept No. 1, for suppose we have collcected responses made by a set of persons to items i and j, and suppose that

(12) f(i,j) of them got item i correct and item j incorrect; and

(13) f(j,k) of them got item j correct and item i incorrect, we note by

(14) g(i,j) = f(i,j) + f(j,i) the number of these persons who receive a total score of 1 on both of the items.

We note, additionally, that

(15) $\frac{f(i,j)}{g(i,j)}$ estimates $\frac{\varepsilon(i)}{\varepsilon(i) + \varepsilon(j)}$,

(16) f(j,i) estimates $\frac{\varepsilon(j)}{\varepsilon(i) + \varepsilon(j)}$, and

(17) $\frac{f(i,j)}{f(j,i)}$ estimates $\frac{\varepsilon(i)}{\varepsilon(j)}$, further substantiating Concept No. 1.

Eqn (17) was derived by Rasch (1970), Choppin (1968), and others, but none, to my knowledge, have used this fact to develop an $\varepsilon(i)$ estimation method, based on the frequency ratios. This process known as the Frequency Ratio Method (FRM) represents the author's modest "contribution" to objective measurement and forms a computational basis for the COMET system. Exhibit 2 shows some available item easiness computational methods.

To continue our derivations consider an item, i, responded to by two persons with different competencies, $\gamma(m)$ and $\gamma(n)$. Again there are 4 outcomes and their probabilities are listed below:

(18) Let $D(m) = 1 + \varepsilon(i) \cdot \gamma(m)$ and $D(n) = 1 + \varepsilon(i) \cdot \gamma(n)$ be the denominator terms in the probability expressions

$$(19) \quad P[X(m,i) = 0,\ X(n,i) = 0] = \frac{1}{D(m) \cdot D(n)}$$

$$(20) \quad P[X(m,i) = 1,\ X(n,i) = 0] = \frac{\varepsilon(i) \cdot \gamma(m)}{D(m) \cdot D(n)}$$

$$(21) \quad P[X(m,i) = 0,\ X(n,i) = 1] = \frac{\varepsilon(i) \cdot \gamma(n)}{D(m) \cdot D(n)}$$

$$(22) \quad P[X(m,i) = 1,\ X(n,i) = 1] = \frac{\varepsilon^2(i) \cdot \gamma(m) \cdot \gamma(n)}{D(m) \cdot D(n)}$$

The conditional probability that the ith item was correctly responded to by the mth person given that the total score made by the mth and nth persons to that item was equal to 1 is given by the equation ratio (24) = (20)/(23) where (23) is given by the equation sum (23)=(20)+(21), thus

$$(23) \quad P[X(m,i) + X(n,i) = 1] = \frac{\varepsilon(i) \cdot \{\gamma(m) + \gamma(n)\}}{D(m) \cdot D(n)}, \text{ and}$$

$$(24)\qquad \frac{(20)}{(23)} = \frac{\gamma(m)}{\gamma(m) + \gamma(n)}$$

which indicates the truth of Concept No. 2, for suppose we have collected responses made to a set of items by persons m and n and suppose

(25) $F(m,n)$ of the items were correctly responded to by person m and were incorrectly responded to by person n, and

(26) $F(n,m)$ of the items were correctly responded to by person n and were incorrectly responded to by person m. We also let

(27) $G(m,n) = F(m,n) + F(n,m)$ be the number of items which were responded to correctly by persons m or n.

We can see immediately that

(28) $\frac{F(m,n)}{G(m,n)}$ estimates $\frac{\gamma(m)}{\gamma(m) + \gamma(n)}$, and

(29) $\frac{F(n,m)}{G(m,n)}$ estimates $\frac{\gamma(n)}{\gamma(m) + \gamma(n)}$, whereas

(30) $\frac{F(m,n)}{F(n,m)}$ estimates $\frac{\gamma(m)}{\gamma(n)}$, further substantiating Concept No. 2.

Rasch (1970) indicates that (30) cannot be practically used to estimate $\gamma(n)$ since the number of items used in an assessment program is usually much smaller than the number of testees. Consequently we will use another approach to estimate $\gamma(n)$ in the FRM or the measurement model analysis sub-system of COMET, although, it seems clear, that in some assessment situations, the number of items is sufficient if the number of $\gamma(n)$ to be estimated corresponds to the number of possible raw scores.

5. The Frequency Ratio Method (FRM)

The FRM is a mathematical-computational procedure designed to provide a simple estimation and goodness of fit test for the BMM parameters and responses. It is intended to be an analytical component of the COMET system. To illustrate FRM we assume we have collected from trainee-persons their responses to the available items of an item bank constructed from the learning (training) objectives of the training course and assumed to indicate training competency if the items are correctly responded to by the trainees. Our available data can be portrayed as the frequencies of equations (12) and (13) and arrayed in a square matrix R(i,j) whose main diagonal is left blank.

After portraying the mathematics and statistics of the FRM we will formulate the computations in two stepwise phases.

Phase I: Item easiness estimation computations.

Phase II: Goodness of Fit computations.

In R(i,j) we form ratios of symmetrical elements and form the square matrix S(i,j). Thus

$$(31)\qquad R(i,j) = \frac{f(i,j)}{f(j,i)} \quad \text{for} \quad i>j,\ i=j=1,\ \ldots,\ I,$$

$$(32)\qquad R(j,i) = \frac{f(j,i)}{f(i,j)} \quad \text{for} \quad j>i,\ j=i=1,\ \ldots,\ I, \text{ and}$$

$$(33)\qquad R(i,j) = \text{blank for } i=1,\ \ldots,\ I.$$

We know from (17) that (31) estimates $\varepsilon(i)/\varepsilon(j)$ and (32) estimates $\varepsilon(j)/\varepsilon(i)$ so that the natural logarithm of (31) is

(34) $\ln S(i,j) = \ln \varepsilon(i) - \ln \varepsilon(j)$ whereas for (32) we have

(35) $\ln S(j,i) = \ln \varepsilon(j) - \ln \varepsilon(i) = -\ln S(i,j)$

If we sum the j^{th} row of $\ln S(i,j)$ we get, after adding and subtracting $\varepsilon(j)$,

(36) $T(j) = \sum_{i=1}^{I} \ln \varepsilon(i) - I \ln \varepsilon(j)$ and let

(37) $\sum_{i=1}^{I} \ln \varepsilon(i) = 0$ for definiteness; and

(38) $\ln \varepsilon(j) = \frac{-T(j)}{I} = t(j)$, and $j = 1, \ldots, I$, so that

(39) $\hat{\varepsilon}(j) = e^{t(j)}$

A stepwise computational procedure is indicated in Exhibit 3. This simple estimation procedure requires only the development and maintenence of a square matrix of frequencies whose elements are the numbers of those persons tested on some of the items and who responded correctly to one, and only one, of each pair of items presented to them. A list of item easiness estimation procedures are shown in Exhibit 2.

Eqn (39) is the culmination of the item easiness estimation phase of the FRM. We now proceed to develop the second phase which is intended to provide a goodness of fit test of the response data to the BMM.

With (39) we are provided with estimates, $\hat{\varepsilon}(i)$, of the easiness parameters, $\varepsilon(i)$. We intend to use the χ^2 distribution (Fisher (1948), as the basis of our goodness of fit test. The hypothesis tested is that the items used for assessment of competency evoke responses in accordance with the BMM. The chi-square form we use is

(40) $\chi^2(i,j) = [f(i,j) - e(i,j)]^2/e(ij)$.

using (12) and (13) for the f(i,j) observed frequencies and substantiating (39) in (15) we define the expected frequencies as

(41) $$e(i,j) = g(i,j) \cdot \frac{\hat{\varepsilon}(i)}{\hat{\varepsilon}(i) + \hat{\varepsilon}(j)}, \quad i>j, \text{ and}$$

(42) $$e(j,i) = g(i,j) \cdot \frac{\hat{\varepsilon}(j)}{\hat{\varepsilon}(i) + \hat{\varepsilon}(j)}, \quad j>i \quad .$$

We can compute χ^2 contributions to the upper and lower parts with the following formalae

(43) $$\chi^2(i,j) = [f(i,j) - e(i,j)]^2/e(i,j), \quad i>j, \text{ and}$$

(44) $$\chi^2(j,i) = [f(j,i) - e(j,i)]^2/e(j,i), \quad j>i \quad .$$

We next add these values symmetrically into a single triangular array by the formula

(45) $$C(i,j) = \chi^2(i,j) + \chi^2(j,i) \quad ,$$

considered as an array with a total of I(I-2) degrees of freedom, so that were we to compute a $\chi^2(k)$ for each item from this table, it would have I-2 degrees of freedom. We compute $\chi^2(k)$ from C(i,j) by summing elements in an "L-shaped" fashion according to the following formula

(46) $$\chi^2(k) = \sum_{j=1}^{k-1} C(i,k) + \sum_{i=k+1}^{I} C(k,j), \quad k=1, \ldots, I \quad .$$

If any $\chi^2(k)$ exceeds the tabled χ^2 for I-2 degrees of freedom at the selected α level, the hypothesis that the responses to the item fit the BMM is rejected and the item should be removed from the COMET system.

Exhibit 4 shows the stepwise computations for these FRM (phase II) goodness of fit calculations.

6. A Numerical Example (Phase 1)

Step 0 This illustration is intended to show the numerical aspects resulting from applying the FRM to a set of response data. The data originated from the responses of Naval recruits to the first I=7 items of the Electronic Technician Selection Test (ETST). Originally 9 items were used but the 8th and 9th item responses were found not to fit the BMM model so they were omitted from the analysis. All recruits responded to the same 7 items so the data does not result from the "catch as catch can" response sampling procedure mentioned earlier. This difference causes no problem in our analysis.

1 The S(i,j) matrix of frequencies

ITEM INCORRECT (columns); ITEM CORRECT (rows)

	1	2	3	4	5	6	7
1	-	220	210	97	203	202	269
2	118	-	157	87	163	148	198
3	111	160	-	62	164	147	205
4	183	275	248	-	277	247	330
5	85	147	145	73	-	130	202
6	164	212	208	123	210	-	245
7	113	144	148	88	164	127	-

2 Compute Frequency Ratios R(i,j)

	1	2	3	4	5	6	7
1	--	1.864	1.892	.530	2.388	1.232	2.381
2	.536	-	.981	.316	1.109	.698	1.375
3	.529	1.019	-	.250	1.131	.707	1.385
4	.887	3.161	4.000	-	3.795	2.008	3.750
5	.419	.902	.884	.264	-	.619	1.232
6	.812	1.432	1.415	.498	1.615	-	1.929
7	.420	.727	.722	.267	.812	.518	-

Step

3 $T(i,j) = \ln R(i,j)$

	1	2	3	4	5	6	7
1	-	.622	.638	-.635	.870	.208	.867
2	-.622	-	-.019	-1.151	.103	-.359	.318
3	-.638	.019	-	1.386	.123	-.347	.326
4	.635	1.151	1.386	-	1.331	.697	1.321
5	-.870	-.103	-.123	-1.331	-	-.480	.208
6	-.208	.359	.347	-.697	.480	-	.657
7	-.867	-.318	-.326	-1.321	-.208	-.657	-

4,5,6 Compute $T(j)$, $t(i)$ and $\hat{\varepsilon}(j)$

	T(j)	t(j)	$\hat{\varepsilon}(j)$
1	2.570	-.367	.693
2	-1.730	.247	1.280
3	-1.903	.272	1.313
4	6.521	-.932	.394
5	-2.699	.386	1.471
6	.438	-.134	.874
7	-3.697	.528	1.696

This completes Phase 1.

Phase 2 - Goodness of Fit Test

Step

0 the $\hat{\varepsilon}(i)$ values are given above in Step 6 of Phase 1

1 the S(ij) matrix, containing f(i,j) and f(j,i), is given in Step 1 of Phase (1). The g(i,j) matrix is given below

	1	2	3	4	5	6	7
1	--	338	321	280	288	366	782
2		--	317	362	310	360	342
3			--	310	309	355	353
4				--	350	370	418
5	(Symmetrical)				--	340	366
6						--	372
7							--

Step

2 Expected Frequencies

	1	2	3	4	5	6	7
1		219.29	210.09	101.50	195.78	204.15	271.18
2	118.71	--	160.53	85.21	165.76	146.09	194.91
3	110.91	156.47	--	71.55	163.28	141.86	198.95
4	178.50	276.79	238.45	--	276.05	255.04	339.21
5	92.23	143.31	145.72	73.94	--	126.72	195.99
6	161.85	213.91	213.14	114.95	213.28	--	245.48
7	110.82	147.09	154.05	78.79	170.01	126.52	--

3 χ^2 Contributions

	1	2	3	4	5	6	7
1	--	.002	.000	.200	.266	.023	.018
2	.004	--	.078	.038	.046	.025	.049
3	.000	.080	--	1.275	.003	.186	.184
4	.113	.112	.382	--	.003	.253	.250
5	.565	.053	.004	.012	--	.085	.184
6	.029	.017	.124	.562	.050	--	.000
7	.043	.065	.238	1.077	.212	.002	--

4 Accumulated χ^2 values

	1	2	3	4	5	6	7
1	--	.006	.000	.313	.831	.052	.061
2		--	.158	.050	.099	.042	.114
3			--	1.657	.007	.031	.422
4				--	.015	.815	1.327
5					--	.002	1.077
6						--	2.322
7							--

5 $\alpha = .05$, $df = I-2 = 5$ $\chi^2(.95, 5) = 11.070 = \chi^2_o$.

Step

6 Compute $\chi^2(k)$

k	$\chi^2(k)$
1	1.263
2	.469
3	2.275
4	4.177
5	2.031
6	2.019
7	5.323

7 No $\chi^2(k)$ exceeds or equals χ_o^2, therefore H_o is accepted for all 7 items.

8 As no items are removed, it is not necessary to re-estimate the $\hat{\varepsilon}(k)$.

This completes Phase 2.

EXHIBIT 1

Types of Measurement Models Developed by Rasch (1960)

Type	Name	Application Fields
CD	Cognitive Domain	Measurement of Intelligence, Aptitudes, Competencies and Abilities
AD	Affective Domain	Measurement of Attitudes, Values Motivation and Interests
SP	Stochastic Process	Measurement involving stochastic process observations such as reading errors

EXHIBIT 2

Item Easiness Estimation Procedures

Method Name	Originator
Least Squares	Bramble (1969)
Graphic	Rasch (1960)
Maximum Likelihood	Wright and Panchapakesan (1969)
Logarithmic	Blommers (1965)
Frequency Ratio	Moonan (1974)

EXHIBIT 3

Stepwise Procedure for FRM (Phase I)
Item Easiness Estimation Computations

Step

0 Collect responses to the items in the item competency bank.

1 Organize the responses into a square matrix S(i,j) whose elements are defined by (12), (13) and (14).

2 Compute the frequency ratios as indicated by (31) and (32). Call the matrix of ratios R(i,j).

3 Take the natural logarithms of the R(i,j) elements; called T(i,j).

4 Sum each row of the matrix of logarithms and store in a vector T(j).

5 Compute $t(j) = -T(j)/(I-1)$

6 Determine the easiness estimates by the relation $\hat{\varepsilon}(j) = e^{t(j)}$.

EXHIBIT 4

Stepwise Procedure for FRM (Phase II)
Goodness of Fit Computation

Step

0 Given the easiness estimates $\hat{\varepsilon}(i)$ of the items whose response fit is desired to be tested.

1 Specify the f(i,j) observed frequencies as defined by eqns (12) and (13) for those items as well as g(i,j) of eqn (14).

2 Compute the expected frequencies as given by (41) and (42).

3 Compute the χ^2 contribution as defined by (43) and (44).

4 Calculate the accumulated contributions of (45).

5 Select an α level for the significance test and find $\chi^2(1-\alpha, I-2) = \chi_0^2$.

6 Compute the test statistics, $\chi^2(k)$, according to (46).

7 If $\chi^2(k) \geq \chi_0^2$ for any k, reject H_0 for those items and remove them from the analysis.

8 If any item is removed, re-estimate, with FRM (Phase I), the $\hat{\varepsilon}(i)$ of the remaining item and repeat steps 1-8 until no items are removed.

7. Estimation of Competency

We do not consider the estimation of the competency parameter of the BMM to be part of the FRM although it is a part of COMET. This arises because it is usually desired to estimate competency on examination responses resulting from administering tests composed of only a few, not all, of the available items. Since, at the time FRM is utilized these subsets have not been selected, the $\gamma(n)$ estimation process is delayed.

Following Wright and Panchapakesan (1969), Moonan and Covher (1973) have programmed a computer subroutine called "Maxco" to estimate $\gamma(n)$ from knowledge of total scores and the easiness of items on an examination. This process uses the Newton-Raphson iteration method to estimate $\gamma(n)$ corresponding to raw score ranging from 1 to k-1 in a test composed of k items. $\gamma(n)$ for raw scores 0 and K are non-estimable, the test being either too difficult or too easy for the examinee. We do know, however, that the competencies are either less than the competency associated with a raw score of 1 or greater than the competency associated with a raw score of k. The formula iterated in Maxco is

$$(47) \qquad f(\gamma(n)) = \sum_{k=1}^{k} \frac{\gamma(n) \cdot \hat{\varepsilon}(k)}{1 + \gamma(n) \cdot \varepsilon(k)} - r \qquad 0<r<k$$

where r is a raw score on a test composed of k items whose easiness are estimated by $\hat{\varepsilon}(k)$ and $\gamma(n)$ is the comptency of person n who obtained a raw score of r on the k items.

We have used Maxco on the I = K = 7 items of the FRM numerical example giving the results in Exhibit 5.

EXHIBIT 5

Competency Estimates for an Examination of I = K = 7 ETST Items

Raw Score	Competency Estimate
0	non-estimable
1	.153
2	.382
3	.743
4	1.363
5	2.628
6	6.469
7	non-estimable

The standard errors of these estimates are not provided here.

8. The COMET System

The goal of Comet is to provide an integrated, systematic and objective procedure for the assessment of trainees in military programs which is based upon the Rasch Binary Measurement Model.

The entire system is composed of 12 interrelated sub-systems and the objectives of COMET are to identify, design and operationdize these sub-systems and integrate them into an effective systematic whole.

This paper's content is not intended to accomplish all of that but is intended as the first step, viz., the project definition and conceptual design phase. Actually it has done more than that since the development of the FRM, shown here, was a necessary sub-system without which the whole system would not be efficient or even feasible. We intend next to specify the essential sub-systems of COMET and to describe their functions. The following sub-systems have been identified

	A Sub-System For	Brief Title
a.	training needs assessment	Needs
b.	transforming training needs into learning objectives	Objectives
c.	training course design	Course
d.	developing test items from learning objectives	Items
e.	item response sampling	Sample
f.	data collection	Data
g.	data analysis	FRM
h.	examination design	Test
i.	test reproduction	Repro
j.	competency estimation	Meas
k.	evaluation	Eval
l.	control	Cont

Description of Sub-System Functions

Needs

Planning the COMET system should start with the identification of training needs. A training need is defined as the discrepency between desired and required outcomes. The important notion is that to have a "need" we must identify and document that there is a gap between two outcomes, that which is currently resulting and that which should be resulting. A training specialist is probably required to design and operate this sub-system.

Objectives

It is convenient and perhaps necessary to transform identified needs into learning objective statements. There are several current means for doing this but none which the author cares to recommend. The training specialist should also be helpful in this sub-system.

Course

As was previously suggested, a training course design procedure needs to be followed to implement this sub-system. Probably no better source for this effort has been prepared other than that of Rundquist (1967).

Items

Having designed the training course and specified learning objectives we need to prepare items which, if answered correctly, will reflect trainee competency. This process may be significantly shortened by utilizing items made available from different services suggested by Wood (1968). Moonan (1972) provides a table of non-military behavioral objective and test item pools.

Sample

Having constructed the items for measuring competency we are required next to administer these items, perhaps in small batches, to individuals of varying training competency. As indicated earlier no formal sampling plan is required but responses may be acquired on a so-called "catch as catch can" basis.

Data

These response data may be recorded in various ways such as port-a-punch or mark sense cards or by an optical scanner form. An optical scanner system together with its mini-computer appears to be the optimal data collection and processing system for COMET. The cost effectiveness of this suggestion needs to be examined in any particular application.

FRM

Having collected the response data it needs to be organized in S(i,j) form and analyzed with the frequency ratio method, FRM in order to estimate the item easiness and to test the goodness of fit of the response data to the BMM. Items not fitting the model should be examined and purged from the item bank.

Test and Repro

A sub-system which can design automatically, prepare and duplicate competency examinations from the available item pool is required. This may be done via the computer or with the clerical system used by GLAKES. This process is sometimes referred to as CATC: Computer Assisted Test Construction.

Meas

Having specified the examination and the items therein we require the development of competency tables associated with possible raw scores on these tests. This is accomplished by using the Maxco subroutine developed by Moonan and Covher (1973).

Eval

Having made the measurements of training competency we can use this information in a variety of ways to make evaluations among which are the following:

a. Trainee competency

b. Change in trainee competency

c. Instructor capabilities

d. Training facility effectiveness

e. Training program effectiveness

f. Program objectives

A good discussion of these and other ways that the BMM can be used in evaluation is given in Hiscox (1974).

Cont

The COMET system needs to be controlled as a system by an effective manager who must constantly carry out the six functions of a manager: planning, directing, organizing, staffing, evaluating and controlling.

Any operational COMET system needs to be developed through systems analysis by an experienced or professional systems analyst.

9. Expected Implementation Problems

It is expected that the most difficult problem to be encountered in implementing COMET would be to find and utilize a systems' analyst to design the system for local conditions and constraints as well as to appoint a competent manager to monitor and manage the operational system. This person must possess many skills and have knowledge in the areas of measurement, computers, education, training, systems, etc. as well as the ability to work with specialists in those fields and with military personnel and systems. Such competent persons are difficult to locate!

Developing or finding items for the competency pool may be onerous but not too difficult. Interservice cooperation should be sought here and the MTA could play a significant and helpful role in these matters. A prototype COMET system might be designed and developed within one service and developmental costs shared. This system might then be used by other interested parties.

Standard errors of estimates of $\gamma(n)$ and $\varepsilon(i)$ are needed but not provided in this report. Their specification is under development but was not completed in time for inclusion in this paper. The science of examination design needs better formulation so that the test sub-system may make use of it. There appears to be only crude rules of thumb and theory available in psychometrics to assist us in this area.

All in all the practical problems do not appear too limiting, what is most needed is the appearance of a dynamic, resourceful, and persuasive leader to step forth from among our ranks and pursue the development and implementation of a COMET-like system within a training facility of his own service.

The major conceptual design work has been shown here, but there are many details yet to be worked out.

10. Summary

This report proposed, defined the scope and outlined the 12 subsystems of a general system called COMET whose general purpose is to objectively assess the competencies attained by trainees during the course of training in the military service. The measurement and evaluation features of COMET are based on a binary measurement model created by George Rasch. A new analysis procedure was developed to estimate a parameter of the model and to test the goodness of fit of responses of test items to the model. A numerical example of computations was provided and practical problems expected to be encountered during implementation are mentioned.

11. References

Blakely, J. A. (1969). Revision of ATAC System; Request for Proposal, Marine Corps Recruit Depot, San Diego.

Blommers, P. (1965). "A ratio Scale for Test Item Difficulty." Unpublished, University of Iowa, Department of Education.

Bramble, W. (1969). "A Least Square Method of Parameter Estimation for the Logistic Measurement Model." Unpublished, University of Kentucky.

Choppin, B. (1968). "Item Bank Using Sample-Free Calibration," Nature, August 24, p. 1-7.

Hatch, R. S. (1967). "Analysis of Automated Testing and Attrition Control Reports," Communication-Electronics Schools, MCRD, San Diego, CA.

Hiscox, S. B. (1974). "The Rasch Model in Evaluation," Paper presented at the Annual Meeting of NCME, Chicago, April, 1974.

Moonan, W. J. (1972). "Evaluation Trainee Test Performances with a G. Rasch Measurement Model," Proceedings of the 1972 MTA, San Antonio.

Moonan, W. J. (1972). "Handout for Presentation on Item Banking with the Rasch Model." Unpublished, NPTRL, San Diego, CA.

Moonan, W. J., and Covher, M. (1973). "MAXCO," Tools for Analysis and Decision, FACT DOCUMENT No. 24, NPTRL, San Diego, CA.

Rasch, G. (1960). Prohabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1969). "An Individual Centered Approach to Item Analysis with Two Categories of Answers." Unpublished, 1969 AERA Pre-Session on Rasch Models, Los Angeles, CA.

Rundquist, E. (1967). Course Design and Redesign Manual for Job Training Courses. San Diego, CA: NPTRL, SRR 66-17 (Revised).

Schmidt, W. (1969). "Rasch Model." Unpublished, 1969 AERA Pre-Session on Rasch Models, Los Angeles, CA.

Wheeler, W. H. (1966). "Learning Evaluation and Improvement Program." Unpublished, Electronics Technician Training School, NTC, Great Lakes, Illinois.

Wood, B. (1968). Educational Research, Vol. 10, No. 2.

Wright, B. (1968). Sample-Free Test Calibration and Person Measurement. Proc. of the 1967 Invitational Conference on Testing Problems. Educational Testing Service.

Wright, B. (1969). Lecture Notes. Unpublished, 1969 AERA Pre-Session on Rasch Models, Los Angeles, CA.

Wright, B., and Panchapakesan, N. (1969). "A Procedure for Sample-Free Item Analysis," Educational and Psychological Measurement, Vol. 29.