DOCUMENT RESUME

ED 272 580 TM 860 499

AUTHOR	Tatsuoka, Kikumi K.; Tatsuoka, Maurice M.
TITLE	Diagnosis of Cognitive Errors by Statistical Pattern Recognition Methods.
PUB DATE	Apr 86
NOTE	30p.; Paper presented at the Annual Meeting of the Psychometric Society (50th, Nashville, TN, June 1-4, 1985).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Artificial Intelligence; Bayesian Statistics; Cognitive Development; *Computer Assisted Testing; Equations (Mathematics); *Error Patterns; *Fractions; Hypothesis Testing; Junior High Schools; *Latent Trait Theory; Mathematical Models; Probability; *Problem Solving; *Response Style (Tests)
IDENTIFIERS	PLATO; *Rule Space

ABSTRACT

The rule space model permits measurement of cognitive skill acquisition, diagnosis of cognitive errors, and detection of the strengths and weaknesses of knowledge possessed by individuals. Two ways to classify an individual into his or her most plausible latent state of knowledge include: (1) hypothesis testing--Bayes' decision rules for minimum errors; and (2) bug distribution--how bugs, incorrect rules used to solve problems, are clustered and related. A 40-item test containing subtraction of fraction items was given to 535 junior high school students. A computer program was used on the PLATO system to diagnose erroneous rules of operation. Two common erroneous problem-solving rules were used to illustrate the rule space model. The results were then compared with the results obtained from a conventional artificial intelligence approach. (GDC)

Diagnosis of Cognitive Errors

by

Statistical Pattern Recognition Methods

by

Kikumi K. Tatsuoka
University of Illinois at Urbana-Champaign
Computer-based Education Research Laboratory

and

Maurice M. Tatsuoka
University of Illinois at Urbana-Champaign
Department of Educational Psychology

April 1986

Abstract

A model (called the rule space model) which permits measuring cognitive skill acquisition, diagnosing cognitive errors, detecting the weaknesses and strengths of knowledge possessed by individuals was introduced earlier. This study further discusses the theoretical foundation of the model by introducing "bug distribution" and hypothesis testing (Bayes' decision rules for minimum errors) for classifying an individual into his/her most plausible latent state of knowledge. The model is illustrated with the domain of fraction arithmetic and compared with the results obtained from a conventional artificial intelligence approach.

Acknowledgements

The authors would like to acknowledge Mr. Robert Baillie for developing several computer programs and for useful discussions concerning this research.

This research was sponsored by the Personnel and Training Research Program, Psychological Sciences Division, Office of Naval Research.

Some of the analyses presented in this report were performed on the PLATO® system. The PLATO® system is a development of the University of Illinois and PLATO® is a service mark of the Control Data Corporation.

Introduction

Several deterministic methods commonly used in Artificial Intelligence have been applied to develop problem-solving programs, or error-diagnostic systems. These methods have successfully diagnosed hundreds of erroneous rules of operation in several domains of arithmetic, algebra, and some areas of science. The results of such error analyses have contributed to our current understanding of human thinking and reasoning.

These approaches, however, fail to take the variability of response errors into account, and also depend on a specific model of problem solving. Therefore, they often cannot diagnose responses affected by random errors (sometimes called "slips") or produced by innovative thinking that is not taken into account by the current models. It is very difficult to develop a computer program whose underlying algorithms for solving a problem represents a wide range of individual differences. Yet, when these diagnostic systems are used in educational practice, they must be capable of evaluating any responses on test-items, including inconsistent performances and those yielded by creative thinking. Recent developments in cognitive psychology and science point out that a student keeps testing his/her hypothesis and evaluating it until learning advances. As stated by VanLehn (1983), "If they are unsuccessful in an attempt to apply a procedure to a problem they are not apt to just quit, as a computer program does. Instead they will be inventive, invoking certain general purpose tactics to change their current process state in such a way that they can continue the procedure" (p.10). Birenbaum and Tatsuoka (1986) showed that inconsistent and volatile applications of rules in signed-number arithmetic is a common phenomenon among nonmasters. Since the 1960's psychometricians have developed probabilistic models to measure latent traits.

As stated by Alvord and Macready (1985), two general classes of latent stucture models have been proposed. These classes have been called Continuum models and State models. For the continuum models, trait acquisition is assumed to be continuous in nature..., whereas for state models, trait acquisition is perceived as an 'all-or-none' process." Paulson (1985) extended the line of research in latent state models to explain erroneous rules of operation in signed-number arithmetic in which each rule is treated as a discrete state. Some basic assumptions in the state models are: first, one must decide how many latent classes or states the model has. Secondly, every subject must belong to exactly one of a finite set of latent classes which are mutually exclusive and exhaustive. Despite recent developments in methods of estimating parameters (Goodman, 1975; Paulson, 1985), probabilistic explanation of volatile changes in the applications of rules is very difficult by state model approaches. Moreover, it is extremely difficult to take all students' performances on a test into account in a single model, especially when several different methods are available to solve a given set of problems. Therefore, we need a model that is capable of diagnosing non-systematic cognitive errors and is also capable of evaluating nonconventional problem-solving activities.

Tatsuoka and her associates (Tatsuoka, 1985, 1984b; Tatsuoka & Linn, 1983; Tatsuoka & Tatsuoka, 1983, 1982) have developed such a model called rule space and have successfully applied it to diagnose misconceptions possessed by students in signed-number and fraction arithmetic. The model maps all response patterns into a set of ordered pairs comprising the latent ability variable θ and one of the IRT-based caution indices (ζ) introduced by Tatsuoka (1984b). However, the approach used in their model lacks, somehow, a sound statistical foundation in expressing The simulation study by Tatsuoka and Baillie (1982) showed that the response patterns yielded by [illegible] imperfect applications of a specific erroneous rule of operation in a procedural domain form a cluster around the rule. Moreover, they found empirically that the two random variables, θ and ζ obtained from those response patterns in the cluster follow an approximate multivariate normal distribution. This cluster around a rule is called a "bug distribution" hereafter. The theoretical foundation of this empirical finding will be discussed in this paper. First, a brief description of the probabilistic model introduced in Tatsuoka (1985) will be given. Then the connection of each "bug distribution" to this model will be discussed in conjunction with the theory of statistical pattern classification and recognition.

Distribution of Responses around an Erroneous Rule

The term "rule" is used loosely, without a precise definition. Tatsuoka and Tatsuoka (1986) say "A rule is a description of a set of procedures or operations that one can use in solving a problem in some well-defined procedural domain such as arithmetic, algebra and the like." A right rule(s) is defined as a rule that produces the right answer to every item in a test, but an erroneous rule may fortuitously yield the right answer for some subset of the items. A logical analysis of cognitive tasks--identifying subtasks for solving the problems correctly, investigating possible solution paths and constructing a subtask tree or process network for a well-defined procedural domain--is often an important prerequisite to developing a cognitive error-diagnostic test. However, theoretical foundations of dealing with such relational databases can be found elsewhere Reingold, Nievergelt and Deo, 1979; Lee, 1983), and they are not our main concerns in this paper. So we here assume that a set of erroneous rules or sources of misconceptions one wishes to diagnose is given _a priori_. Indeed it is possible to predict a set of erroneous rules by carrying out a detailed, logical task analysis. (Klein, et al., 1981). Further, we assume that each rule yields its unique response pattern(s) to the test items. (The unit of scoring can be the final answer or subprocessess.) Some rules are combinations of the right rule and wrong rules, while others are combinations of various wrong rules. For example, suppose a 40-item fraction subtraction test contains items requiring borrowing and those that do not. If a student increases the numerator by 10 instead of adding the denominator when borrowing, then his answer will most likely be wrong for the items requiring borrowing but correct for those not requiring borrowing. Therefore, this rule--referred to as Rule 8 later--corresponds to the response pattern of zeros for the borrowing items and ones for the non-borrowing items. The set of rules in a study is by no means a complete list of rules. Indeed, we will show that some responses are impossible to diagnose.

The responses around a particular rule of operation in a procedural domain which are produced by not-perfectly-consistent applications of the rule to the test items form a cluster. They include responses that deviate, in various degrees, from the response generated by the rule. When these discrepancies are observed, they are considered as random errors. These random errors are called "slips" by cognitive scientists (Brown & VanLehn, 1980). The properties of such responses around a given erroneous rule will be investigated in this section.

First, the probability of having a "slip" on item j ($j=1,2,\ldots,n$) is assumed to have the same value, p, for all items and it will be called "slip probability" in this paper. Let us denote an arbitrary rule for which the total score is r by Rule R and let the corresponding response pattern be:

(1) $R = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \\ x_{r+1} \\ \vdots \\ x_n \end{bmatrix}$, $x_1 = x_2 \cdots x_r = 1$, and $x_{r+1} = \cdots = x_n = 0$.

The response patterns existing one slip away from Rule R are of two kinds: a slip of "one to zero" occurring at $1 \le j \le r$ and "zero to one" at $r < j \le n$. The number of response patterns having one slip is therefore $\binom{r}{1}\binom{n-r}{0} + \binom{r}{0}\binom{n-r}{1}$, and the probability of having one slip on item $j (j=1,\ldots,n)$ is given by $\binom{r}{1} p^1 (1-p)^{r-1} \binom{n-r}{0} p^0 (1-p)^{n-r} + \binom{r}{0} p^0 (1-p)^r \binom{n-r}{1} p^1 (1-p)^{n-r-1}$ if the probability p is the same for all items, $j=1,\ldots,n$. Therefore the following equation (2) is obtained:

(2) $\text{Prob } (x_j \ne 1 \text{ for some } j=1,\ldots,r \text{ or } x_j \ne 1 \text{ for some } j=r+1,\ldots,n) =$

$$\text{Prob (having a slip on an item)} = \left\{\binom{r}{1}\binom{n-r}{0} + \binom{r}{0}\binom{n-r}{1}\right\} p^1 (1-p)^{n-1} \quad .$$

Similarly, the probability of having k slips on the items is given by as follows:

Prob (having k slips on the items)

$$= \left\{ \sum_{k_1 + k_2 = k} \binom{t}{k_1} \binom{n-t}{k_2} \right\} p^k (1-p)^{n-k}.$$

The generating function of the distribution of frequencies up to k slips will be given by Equation (3) as follows:

$$(3) \quad \sum_{s \le k} \text{Prob (having up to k slips)} = \sum_{s \le k} \binom{n}{s} p^s (1-p)^{n-s}.$$

Therefore, a cluster around Rule R which consists of response patterns including various numbers of slips (not-perfectly-consistent applications of Rule R) has a probability distribution of the binomial form if all items have the same slip probability p. If, on the other hand, we assume each item to have an unique slip probability, then the binomial distribution expressed by Equation (3) will become a compound binomial distribution, Equation (4).

$$(4) \quad \text{Prob (having up to k slips)} = \sum_{x, s \le k}^{k} \left\{ \prod_{j}^{n} p_j^{x_j} (1-p_j)^{1-x_j} \right\}$$

Before an approximation of the slip probabilities p_j is discussed, the rule-space concept will be briefly introduced in the next section.

A Brief Summary of the Probabilistic Model, Rule Space

One of the purposes of the model, the rule space, is to interpret semantically the relationships among various erroneous rules and the right rule, and compare the characteristics of each rule to the right rule or other rules. An analogy for the underlying motivation of seeking a norm-referenced characteristic of "bug behavior" may be found in the theory and practice of norm-referenced tests. This starts by selecting the right rule as a norm and then comparing the other erroneous rules to the characteristic of the norm. By doing so, the psychometric behavior of "bugs" as compared with the right rule, understanding why and how various misconceptions are related and transformed from one to another will be explained more clearly than by just describing the list of bugs.

The rule space model begins by mapping all possible binary response patterns into a set of ordered pairs $\{(\theta, \zeta)\}$, where θ is the latent ability variable in item response theory (IRT) and ζ (or $\zeta(\underset{\sim}{x} : \zeta)$) is one of the IRT-based caution indices (Tatsuoka, 1984b; Tatsuoka & Linn, 1983). The mapping function $f(\underset{\sim}{x})$ is expressed as an inner product of two residual vectors, $P(\underset{\sim}{\theta}) - \underset{\sim}{x}$ and $P(\underset{\sim}{\theta}) - T(\underset{\sim}{\theta})$ where $P_j(\theta)$, $j=1,\ldots,n$ are the one- or two-parameter logistic-model probabilities, $\underset{\sim}{x}$ is a binary response vector and $\underset{\sim}{T}(\theta)$ is the mean vector of the logistic probabilities. $f(\underset{\sim}{x})$ is a linear mapping function between $\underset{\sim}{x}$ and ζ at a given level of θ, and the response patterns having the same sufficient statistics for the maximum likelihood estimate $\hat{\theta}$ of θ are dispersed into different locations on the line of $\theta = \hat{\theta}$. For example, on a 100-item test, there are 4950 different response patterns having the total score of 2. The ζ's for the 4950 binary patterns will be distributed between ζ_{min} and ζ_{max} where ζ_{min} is obtained from the pattern having 1 for the two easiest items and zeros elsewhere, and ζ_{max} is from the pattern having 1 for the two most difficult items. $f(\underset{\sim}{x})$ has the expectation zero and variance $\sum_{j=1}^{n} P_j(\theta)\, Q_j(\theta)\, (P_j\theta - T(\theta))^2$ (Tatsuoka, 1985). Since the expectation of the random variable x_j $(j=1,\ldots,n)$ is $P_j(\theta)$, the expectation of a vector $\underset{\sim}{x}$ is $P(\underset{\sim}{\theta})$ whose jth component is $P_j(\theta)$. The vector $\underset{\sim}{P}(\theta)$ will be mapped to zero as shown in (5), thus the pattern corresponds to $(\theta, 0)$ in the rule space.

(5) $\quad f(\underset{\sim}{P}(\theta)) = 0$

As for an erroneous rule R, the response vector $\underset{\sim}{R}$ given by (5) will be mapped onto $(\theta_R, f(\underset{\sim}{R}, \theta_R))$, where the f value is $\sum_{j=1}^{n} (P_j(\theta) - R_j)(P_j(\theta) - T(\theta))$, and is given by (6). That is,

$$\text{(6)} \qquad f(\underset{\sim}{R}) = -\sum_{j=1}^{r} Q_j(\theta_R)\big(P_j(\theta_R) - T(\theta_R)\big) + \sum_{j=r+1} P_j(\theta_R)\big(P_j(\theta_R) - T(\theta_R)\big) .$$

Similarly, all the response vectors resulting from several slips around rule R will be mapped in the vicinity of $(\theta_R, f(\underset{\sim}{R}))$ in the rule space and form a cluster (called the cluster around R hereafter).

The two variables $\hat{\theta}$ and $f(\underset{\sim}{x})$ are mutually uncorrelated so their covariance matrix has a diagonal form as follows;

$$\text{(7)} \qquad \begin{bmatrix} \mathrm{var}(\hat{\theta}) & 0 \\ 0 & \mathrm{var}(f(\underset{\sim}{x})) \end{bmatrix} = \begin{bmatrix} 1/I(\hat{\theta}) & 0 \\ 0 & \Sigma P_j(\hat{\theta}) Q_j(\hat{\theta}) \big(P_j(\hat{\theta}) - T(\hat{\theta})\big)^2 \end{bmatrix}$$

where $I(\hat{\theta})$ is the information function of the test and is given appoximately by $\Sigma a_j^{\,2} P_j(\theta) Q_j(\theta)$ where the a_j $(j=1,\ldots,n)$ are item discriminating powers.

Let us map all response patterns of the test, including clusters around various rules into the Cartesian product space of $\hat{\theta}$ and $f(\underset{\sim}{x})$, where

$$\text{(8)} \qquad f(\underset{\sim}{x}) = \big(\underset{\sim}{P}(\theta), \underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta)\big) - \big(\underset{\sim}{x}, \underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta)\big)$$

In particular, Rule R itself will be mapped as

$$\text{(9)} \qquad \underset{\sim}{R} = \underset{\sim}{x} \rightarrow (\hat{\theta}_R, f(\underset{\sim}{R})) ,$$

where $f(\underset{\sim}{R})$ is given by Equation (9). The variance of the cluster around R will be expressed by [illegible]:

(10) Var (the cluster around R) = [illegible] .

The quantities p_j and q_j are associated with Rule R as well as with item j, and their values are unknown.

Slip probabilities

Suppose p_j is the slip probability of item j and $p_j \neq p_l$ for $j \neq l$. Then, the probability density function of a cluster around Rule R will be a compound binomial distribution. The conditional probability that x_j, the response to item j, is not equal to the jth element of Rule R, x_{R_j}, but $1 - x_{R_j}$ will be either $P_j(\theta)$ or $Q_j(\theta)$ depending on whether the jth element x_{R_j} in R is zero or one, respectively. That is

(11) $$\text{Prob}(x_j \neq x_{R_j} \mid \theta_R) = \begin{cases} \text{Prob}(x_j = 1 \mid \theta_R) = P_j(\theta_R) & \text{if } x_{R_j} = 0 \\ \text{Prob}(x_j = 0 \mid \theta_R) = Q_j(\theta_R) & \text{if } x_{R_j} = 1 \end{cases}$$

Therefore, the slip probability of item j will be expressed by the logistic function $P_j(\theta)$ whose parameters are estimated from a sample. The compound binomial distribution of the cluster around Rule R is given by the terms of the expansion of expression (12), and the mean and variance by Equations (13) and (14) because the complement of the slip probability is the conditional probability of correct responses given Rule R.

(12) $$g(R) = \prod_{j=1}^{n} \left(P_j(\theta_R) + Q_j(\theta_R) \right)$$

(13) $$\mu_R = \sum_{j=1}^{r} P_j(\theta_R) + \sum_{j=r+1}^{n} Q_j(\theta_R)$$

(14) $\sigma_R = \sum_{j=1}$ [illegible]

After mapping the distribution function $g(\theta_R)$ of Rule R into the rule space by the mapping function, $f(\zeta)$, the centroid and covariance matrix will be given by equations (15) and (16), respectively.

(15) Var(ζ in the cluster around R) $\simeq \Sigma P_j(\theta) Q_j(\theta) (P_j(\theta) - T(\theta))^2$

The variance of θ in any cluster, on the other hand, is given by the reciprocal $1/I(\theta)$ of the information function

(16) Var(θ in the cluster around R) $= 1/I(\theta_R)$

The above two var ces, along with the fact that ζ and $\hat{\theta}$ are uncorrelated, plus the reasonable assumption that they have a bivariate normal distribution, allow us to construct any desired percent ellipse around each rule point $\underset{\sim}{R}$. The upshot is that, if all erroneous rules (and the correct one) were to be mapped into the rule space along with their neighboring response patterns representing random slips from them, the resulting topography would be something like what is seen in Figure 1. That is, the population of points would exhibit modal densities at many rule points that each forms the center of an enveloping ellipse with the density of points getting rarer as we depart farther from the center in any direction. Furthermore, the major and minor axes of these ellipses would -- by virtue of the uncorrelatedness of ζ and $\hat{\theta}$ -- be parallel to the vertical (ζ) and horizontal ($\hat{\theta}$) reference axes of the rule space, respectively.

Insert Figure 1 about here

R[illegible] and [illegible] multiples of the respective standard deviations

$$\left[\sum_{j=1}^{r} P_j(\theta) Q_j(\theta) \left(P_j(\theta) - T(\theta)\right)^2\right]^{\frac{1}{2}} \quad \text{and} \quad I(\theta)^{-\frac{1}{2}}$$

we may assert that the set of ellipses gives a complete characterization of the rule space. By this is meant that, once these ellipses are given, any response-pattern points can be classified as most likely being a random slip from one or another of the erroneous rules (or the correct one). We have only to determine, for a suitable percent value, which one of the several ellipses uniquely includes the given point.

Operational Classification Scheme

The geometric scheme outlined above for classifying any given response-pattern point as being a "perturbation" from one or another of the rule points has a certain intuitive appeal (especially to those with high spatial ability!). However, it is obviously difficult if not infeasible to put it into practice. We therefore now describe the algebraic equivalent of the foregoing geometric classification decision rule, which is none other than the well-known minimum-D^2 rule, where D^2 is Mahalanobis' generalized squared-distance (Fukunaga, 1972; Tatsuoka, 1971). Then the Bayes' decision rule for minimum error will be discussed in the context of the rule space.

Without loss of generality, we may suppose that a given response pattern point $\underset{\sim}{x}$ has to be classified as representing a random slip from one of two rule points $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$. Let $\underset{\sim}{X}$ be a point in the rule space corresponding to $\underset{\sim}{x}$, $\underset{\sim}{X} = \begin{bmatrix} \theta_x \\ f(x) \end{bmatrix}$. The estimated Mahalanobis distance of $\underset{\sim}{X}$ from each of the two rule points is

$$(17)\quad \hat{D}^2_{x_j} = (\underset{\sim}{X} - \underset{\sim}{R}_j)' \, \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{X} - \underset{\sim}{R}_j) \quad j=1,2$$

where $\underset{\sim}{R}_1 = \begin{bmatrix} \hat{\theta}_{R_1} \\ f(\underset{\sim}{R}_1) \end{bmatrix}$ and $\underset{\sim}{R}_2 = \begin{bmatrix} \hat{\theta}_{R_2} \\ f(\underset{\sim}{R}_2) \end{bmatrix}$, and the variance-covariance matrix

will be,

$$\underset{\sim}{\Sigma} = \begin{bmatrix} 1/I(\hat{\theta}) & 0 \\ 0 & \mathrm{var}\left(f(\underset{\sim}{x})\right) \end{bmatrix}$$

The decision rule is, of course, to classify $\underset{\sim}{x}$ as a perturbation from $\underset{\sim}{R}_1$ if $\hat{D}^2_{x1} < \hat{D}^2_{x2}$ and otherwise as a perturbation from $\underset{\sim}{R}_2$. However, the decision based on the Mahalanobis distances, $\hat{D}^2_{x1}$ and $\hat{D}^2_{x2}$ does not provide error probabilities of misclassification. The next section will discuss them.

<u>The Bayes' Decision Rule for Minumum Error</u>

Suppose R_1 and R_2 are two clusters of points corresponding to Rules 1 and 2, respectively.

Let Prob (R_1) and Prob (R_2) be prior probabilities of the rules R_1 and R_2, $p(\underset{\sim}{Y} \mid R_i)$, $i=1,2$ be the conditional density function of $\underset{\sim}{Y}$ given R_i. Then, Bayes' decision rule is as summarized in Equation (18).

$$(18)\qquad \text{If } p(\underset{\sim}{Y} \mid R_1)\ \mathrm{Prob}(R_1) > p(\underset{\sim}{Y} \mid R_2\ \mathrm{Prob}(R_2) \text{ then } \underset{\sim}{Y} \in R_1$$

$$\text{Otherwise, } \underset{\sim}{Y} \in R_2 \quad .$$

Sometimes, it is convenient to take the negative log of the likelihood ratio in Expression (18) and rewrite it as Expression (19).

(19) $$\text{If } h(\underset{\sim}{Y}) = \ln l(\underset{\sim}{Y}) = -\ln\left(p(\underset{\sim}{Y} \mid R_1)\right) + \ln\left(p(\underset{\sim}{Y} \mid R_2)\right)$$

$$\ln\left[\text{Prob}(R_1) / \text{Prob}(R_2)\right] \text{ then } \underset{\sim}{Y} \text{ belongs to } R_1 .$$

The probability of error is the probability that $\underset{\sim}{Y}$ will be assigned to the wrong group, R_1.

Let us denote the posterior density function by $P(\underset{\sim}{R}_1 \mid \underset{\sim}{Y})$ and let Γ_1 and Γ_2 be the regions such that if $\underset{\sim}{Y} \in \Gamma_1$ then $P(R_1 \mid \underset{\sim}{Y}) > P(R_2 \mid \underset{\sim}{Y})$ and if $\underset{\sim}{Y} \in \Gamma_2$ then $P(R_2 \mid \underset{\sim}{Y}) > P(R_1 \mid \underset{\sim}{Y})$.

The probability of error is given by the following equation:

(20) $$\epsilon = \text{Prob}(\underset{\sim}{Y} \in \Gamma_2 \mid R_1)\, P(R_1) + \text{Prob}(\underset{\sim}{Y} \in \Gamma_1 \mid R_2)\, P(R_2) .$$

Let us denote the probability of $\underset{\sim}{Y}$ belonging to Γ_2 when $\underset{\sim}{Y}$ is from R_1 by ϵ_1, then

(21) $$\epsilon_1 = \text{Prob}(\underset{\sim}{Y} \in \Gamma_2 \mid R_1) = \int_{\Gamma_2} p(\underset{\sim}{Y} \mid R_1)\, dY,$$

Similarly, the probability of $\underset{\sim}{Y}$ belonging to Γ_1 when $\underset{\sim}{Y}$ is from R_2, ϵ_2 will be

(22) $$\epsilon_2 = \text{Prob}(Y \in \Gamma_1 \mid R_2) = \int_{\Gamma_1} p(\underset{\sim}{Y} \mid R_2)\, d\underset{\sim}{Y} .$$

Then expression (20) can be rewritten as $\epsilon = \epsilon_1 P(R_1) + \epsilon_2 P(R_2)$, or more precisely,

(23) $$\epsilon = P(R_1)\int_{\Gamma_2} p(\underset{\sim}{Y} \mid R_1)\, d\underset{\sim}{Y} + P(R_2)\int_{\Gamma_1} p(\underset{\sim}{Y} \mid R_2)\, d\underset{\sim}{Y} .$$

That is, the total probability of error is a weighted [illegible] classification of samples from R_1 and R_2 into R_2 and R_1, respectively.

The integration of the conditional density function is necessary to get the error probability ϵ. The dimensionality of the conditional density function is often more than one, while the density function $p(\ell \mid R_1)$ of the likelihood ratio is one dimensional so it is sometimes convenient to integrate the latter (Fukunaga, 1972). Hence, Equations (24) and (25) are used to obtain the error probabilities, ϵ_1 and ϵ_2;

(24) $$\epsilon_1 = \int_0^{P(R_2)/P(R_1)} p(\ell \mid R_1)\, d\ell$$

(25) $$\epsilon_2 = \int_{P(R_2)/P(R_1)}^{\infty} p(\ell \mid R_2)\, d\ell$$

If the density function $p(\underset{\sim}{Y} \mid R_i)$ is normal with expectations $\underset{\sim}{M}_i$ and covariance matrices $\underset{\sim}{\Sigma}_i$, then Equation (19) will become Equation (26).

(26) If $h(\underset{\sim}{Y}) = -\ln \ell(\underset{\sim}{Y})$

$$= \frac{1}{2}(\underset{\sim}{Y} - \underset{\sim}{M}_1)' \underset{\sim}{\Sigma}_1^{-1} (\underset{\sim}{Y} - \underset{\sim}{M}_1) - \frac{1}{2}(\underset{\sim}{Y} - \underset{\sim}{M}_1)' \underset{\sim}{\Sigma}_2^{-1} (\underset{\sim}{Y} - \underset{\sim}{M}_2)$$

$$+ \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} \begin{matrix} < \\ > \end{matrix} \ln \frac{P(R_1)}{P(R_2)} \rightarrow \begin{cases} Y \in R_1 \\ Y \in R_2 \end{cases} .$$

If $\underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2 = \underset{\sim}{\Sigma}$, then $h(\underset{\sim}{Y})$ becomes a linear function of $\underset{\sim}{Y}$ and the decision rule has the following form if $\underset{\sim}{Y}$ follows a normal distribution:

$$h(Y) = \frac{1}{2}(Y - M_1)'\Sigma^{-1}(Y - M_1) - \frac{1}{2}(Y - M_2)'\Sigma^{-1}(Y - M_2)$$

$$= \frac{1}{2}\left\{(M_2' - M_1')\Sigma^{-1}Y - Y'\Sigma^{-1}(M_1 - M_2) + M_1'\Sigma^{-1}M_1 - M_2'\Sigma^{-1}M_2\right\}$$

(27)

$$= (M_2' - M_1')\Sigma^{-1}Y + \frac{1}{2}(M_1'\Sigma^{-1}M_1 - M_2'\Sigma^{-1}M_2) \lessgtr \ln\left[P(R_1)/P(R_2)\right]$$

$$= t \quad .$$

then, $\quad Y \in \begin{cases} R_1 \\ \\ R_2 \end{cases}$

The error probability ϵ_1 is given by,

(28) $$\epsilon_1 = \int_t^{\infty} p\big(h(Y) \mid R_1\big)\, dh(Y) = \int_{\frac{t+\eta}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-Z^2}{2}\right) dZ$$

$$= 1 - \Psi\left(\frac{t+\eta}{\sigma}\right) \quad .$$

where $t = \ln \; p(R_1) \,/\, p(R_2)$ and Ψ (.) is the unit normal distribution. The conditional expectation of the likelihood function $h(Y)$ is given by (29) and (30),

(29) $$E\big(h(Y) \mid R_1\big) = -{}^1/_2\,(M_2 - M_1)'\,\Sigma^{-1}\,(M_2 - M_1) = -\eta$$

(30) $$E\big(h(Y) \mid R_2\big) = +{}^1/_2\,(M_2 - M_1)'\,\Sigma^{-1}\,(M_2 - M_1) = +\eta$$

and the variance of $h(Y)$ is given by Equation (31):

(31) $$\sigma_1 = E[(h(Y) - \eta)^2 \mid R_1] = (M_2 - M_1)'\Sigma^{-1}(M_2 - M_1) = 2\eta \quad .$$

Similarly, ϵ_2 can be obtained by calculating $1 - \Psi(\frac{\eta - t}{\sigma})$, i.e.,

(32) $$\epsilon_2 = \int_{-\infty}^{t} p(h(Y) \mid R_2)\, dh(Y) = 1 - \Psi\left(\frac{\eta - t}{\sigma}\right) \quad .$$

<u>Illustration of the model with an example</u>

A 40-item fraction subtraction test was given to 535 students at a local junior high school. A computer program adopting a deterministic strategy for diagnosing erroneous rules of operation in subtracting two fractions was developed on the PLATO system. The students' performances on the test were analyzed by the error-diagnostic program and summarized by Tatsuoka (1984b). In order to illustrate the rule space model and the decision rule described in the previous section, two very common erroneous rules (Tatsuoka, 1984b) are chosen to explain the model.

Rule 8. This rule is applicable to any fraction or mixed number. A student subtracts the smaller from the larger number in unequal corresponding parts and keeps corresponding equal parts as is in the answer. Examples are,

1. $4\frac{4}{12} - 2\frac{7}{12} = 2\frac{3}{12} = 2\frac{1}{4}$
2. $7\frac{3}{5} - \frac{4}{5} = 7\frac{1}{5}$
3. $\frac{3}{4} - \frac{3}{8} = \frac{3}{4}$

Rule 30. This rule is applicable to the subtraction of mixed numbers where the first numerator is smaller than the second numerator. A student reduces the whole-number part of the minuend by one and adds one to the tens digit of the numerator.

1. [illegible]

2. $3\frac{3}{8}$ [illegible] $= 2\frac{13}{8}$ [illegible]

3. $7\frac{3}{5} - \frac{4}{5} = 6\frac{13}{5} - \frac{4}{5} = 2\frac{9}{5}$

These two rules are applied to 40 items and two sets of responses are scored by the "right or wrong" scoring procedure. The binary score pattern made by Rule 8 is denoted by R_8 and the other made by Rule 30 is denoted by R_{30}.

Besides the two rules mentioned above, 38 different error types are identified by a task analysis. However, these error types do not necessarily represent microlevels of cognitive processes such as erroneous rules of operation. They are, somehow, defined more coarsely, like borrowing errors being grouped as a single error type, or the combination of borrowing and getting the least common multiple of two denominators being counted as one error type. In other words, 38 binary response patterns representing 38 error types are obtained.

The 535 students' responses on the 40 items are scored and used for estimating item parameters a_j and b_j by the maximum likelihood procedure. By using these a- and b-values, θ-values associated with the two rules and 38 error types are computed. Then the corresponding ζ-values are calculated. Thus, 40 points, $(\hat{\theta}_k, \zeta_k)$, k=1,...,40 are plotted in the rule space (Rule 8 is renumbered to 39 and Rule 30 to 40. It is only a coincidence that the number of rules equals the rule number.)

Insert Table 1 about here

Now, two students A and B who used Rules 8 and 30 for a subset of 40 items are selected. This was possible because their performances are diagnosed independently by the error diagnostic system SPFBUG mentioned in Tatsuoka (1984). The circles shown in Figure 2 represent A and B. Their estimated Mahalanobis distances, $\hat{D}^2$, to the 40 centroids are calculated respectively and the smallest values of two distances, $\hat{D}^2$, are selected to compute probabilities of errors. Table 2 summarizes the results.

Insert Table 2 & Figure 2 about here

The $\hat{D}^2$ values of Student A to Sets 40 and 19 are 0.008 and 0.119, respectively, and both the values are small enough to judge that A may be classified to either of the sets. Since $\hat{D}^2$ follows the χ^2-distribution with two degrees of freedom (Tatsuoka, 1971) the null hypotheses that $D^2_{(A,\text{Set }40)} \equiv 0$ and $D^2_{(A,\text{Set }19)} \equiv 0$ cannot be rejected at, say $\alpha = .25$. The error probabilites ϵ_1 and ϵ_2 are .581 and .266, respectively. Therefore, we conclude A belongs to Set 19, even though $D^2_{(A,\text{Set }40)}$ is smaller than $D^2_{(A,\text{Set }19)}$. This is because the prior probability of Prob (Set 40) is much smaller than that of Prob (Set 19) where the threshold value, t, is determined as follows:

$$t = -\ln\left[\text{Prob (Set 40)} \;/\text{Prob (Set 19)}\right]$$

and

$$\text{Prob (Set k)} \propto (1/2\pi)\exp\left[-(\hat{\theta}_k,\zeta_k)'\,\Sigma_k^{-1}\,(\hat{\theta}_k,\zeta_k)/2\right] .$$

Discussion

A new probabilistic model that is capable of measuring cognitive-skill acquisition, and of diagnosing erroneous rules of operation in a procedural domain was introduced by Tatsuoka and her associates (Tatsuoka, 1985; Tatsuoka & Baillie, 1983; Tatsuoka & Tatsuoka, 1982; Tatsuoka, 1983; Tatsuoka, 1984a). The model, called rule space, involves two important components: 1) determination of a set of rules to be diagnosed, or in other words, conditional density functions representing clusters around the rules, and 2) establishment of decision rules for classifying an observed response pattern into one of the clusters around the rules and computing error probabilities. If each cluster around a rule can be described by a bivariate normal distribution of θ and ζ, the application of the techniques available in the theory of statistical classification and pattern recognition is fairly straightforward. With regards to the first component, a list of rules is supplied independently from parameter estimation of the Item Response Theory models. Diagnoses of students' responses to the items are performed by classifying them into one of the bug distributions if possible, and if not possible then left for the future investigation as to searching a cause of misclassification. Determination of the list of the rules will be discussed in a future paper.

This study introduces the fact that the cluster around the rule consisting of the response patterns resulting from one, two,..., several slips away from perfect application of the rule indeed follows a compound binomial distribution with centroid (θ_R, ζ_R) and variance $\sum_{j=1}^{n} p_j q_j$, where p_j ($j=1,\ldots,n$ is the probability of having a slip from Rule R for item j. The values of p_j and q_j are replaced by the logistic probabilites $P_j(\theta_R)$ and $Q_j(\theta_R)$, $j=1,\ldots,n$, estimated from the dataset.

Appropriateness of using the slip probabilities associated with each erroneous rule by the logistic function is left as a future topic of investigation, although the fit with the data seems to be good.

The determination of a set of ellipses representing clusters around the rules can be automatic after all the erroneous rules are discovered. Many researchers in cognitive science and artificial intelligence have started constructing error diagnostic systems in various domains in this decade. Expert teachers usually know their students' errors, as well as the weaknesses and strengths of each child's knowledge structure. Since the model does not require a large-scale computation such as strategies commonly used in the area of artificial intelligence do, the rule-space model is helpful in more general areas of research and teaching, and for those who have microcomputers for testing their hypotheses, validating their data with probabilistically sound information, and evaluating their teaching methods and materials. Moreover, the model can be "intelligent" in the sense that the researcher can improve and modify the information for the cluster ellipses as they get more new students whose performances they can study.

The set of ellipses can represent many things besides erroneous rules. They can represent specific contents of some domain, usage errors in the langauge arts, or processes required in algebra. However, further research is necessary to develop methods for determining the set of ellipses other than relying on an expert teacher. The method must be efficient and compatible with recent theories of human cognition and learning.

References

Alvord, G., & Macready, G. B. (1985). Comparing fit of nonsubsuming probability models. Applied Psychological Measurement, 9, 3, 233-240.

Birenbaum, M., & Tatsuoka, K. K. (1986). On the stability of students' rules of operation for solving arithmetic problems (Technical Report 86-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Brown, J. S., & VanLehn, K. (1980). Repair Theory: A generative theory of bugs in procedural skills. Cognitive Science, 4, 4,

Fukunaga, K. (1972). Introduction to statistical pattern recognition. NY: Academic Press.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. Journal of the American Statistical Association, 70, 755-768.

Klein, M., Birenbaum, M., Standiford, S., & Tatsuoka, K. K. (1981). Logical error analysis and construction of tests to diagnose student "bugs" with addition and subtraction of fractions (Technical Report 81-6). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Lee, T. T. (1983). An algebraic theory of relational databases. The Bell System Technical Journal, 62, 10, 3161-3128.

Paulson, J. A. (1985). Latent class representation of systematic patterns in test responses (ONR research report). Portland: Portland State University.

Reingold, E. M., Nievergelt, J., & Deo, N. (1977). Combinatorial algorithms, theory and practice. Englewood Cliffs, NJ: Prentice-Hall.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 4, 345-354.

Tatsuoka, K. K. (Ed.) (1984a). Analysis of errors in fraction addition and subtraction problems (Final Report for Grant No. NIE-G-81-0002). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, K. K. (1984b). Caution indices based on item response theory. Psychometrika, 9, 95-110.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. Journal of Educational Statistics, 10, 1, 55-73.

Tatsuoka, M. M. (1971). Multivariate analysis: Techniques for educational and psychological research. NY: John Wiley & Sons.

Tatsuoka, K. K., & Baillie, R. (1982). Rule space, the product space of two score components in signed-number subtraction: An approach to dealing with inconsistent use of erroneous rules (Technical Report 82-3-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual response patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7, 1, 81-96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant reponse patterns. Journal of Educational Statistics, 7, 3, 215-231.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 20, 3, 221-230.

Tatsuoka, M. M., & [illegible] [illegible] (Eds.) *Encyclopedia [illegible]*. New York: W[illegible]

Table 1

The 40 Centroids Representing 40 different error types in Fraction Subtraction Tests (N = 535, n = [illegible]0)

Group	θ	ζ	No. of Items	Group	θ	ζ	No. of Items
1	-2.69	-.80	1	21	.24	-.89	22
2	-1.22	-.69	4	22	-.22	-1.23	14
3	-.75	-.68	8	23	.62	-1.55	32
4	-.46	.75	10	24	1.04	-.61	38
5	.11	.91	18	25	.75	-.05	34
6	.64	1.74	30	26	-.51	-1.62	10
7	-.17	1.48	13	27	-.87	-.56	6
8	.40	-.16	25	28	-1.99	1.01	2
9	.60	-.43	31	29	-.19	1.53	12
10	.57	-.24	29	30	-[illegible]4	2.74	10
11	.99	.72	37	31	-1.18	1.46	4
12	1.19	.86	39	32	-1.45	.58	4
13	-.60	-1.58	10	33	.64	1.74	30
14	-.44	-2.31	12	34	.57	-.66	31
15	-.18	.67	14	35	.59	-1.39	30
16	-.08	-1.81	16	36	-1.66	-1.96	4
17	.16	-.86	20	37	-.52	-.94	10
18	-.01	-2.12	18	38	-.32	-1.26	14
19	.09	-2.26	20	39	-.41	-2.57	13
20	.29	-1.51	24	40	.17	-2.34	22

*These items will have the score of 1, otherwise the score will be 0.

Table 2

Summary of Classification Results of Students A and B

	Student A		Student B	
D^2	D^2_{A}, Set 40	.008	D^2_{B}, Set 39	.021
	D^2_{A}, Set 19	.119	D^2_{B}, Set 14	.135
ε_1	.581		.979	
ε_2	.266		.010	
η	.088		.040	
t	-.174		-.61[illegible]	

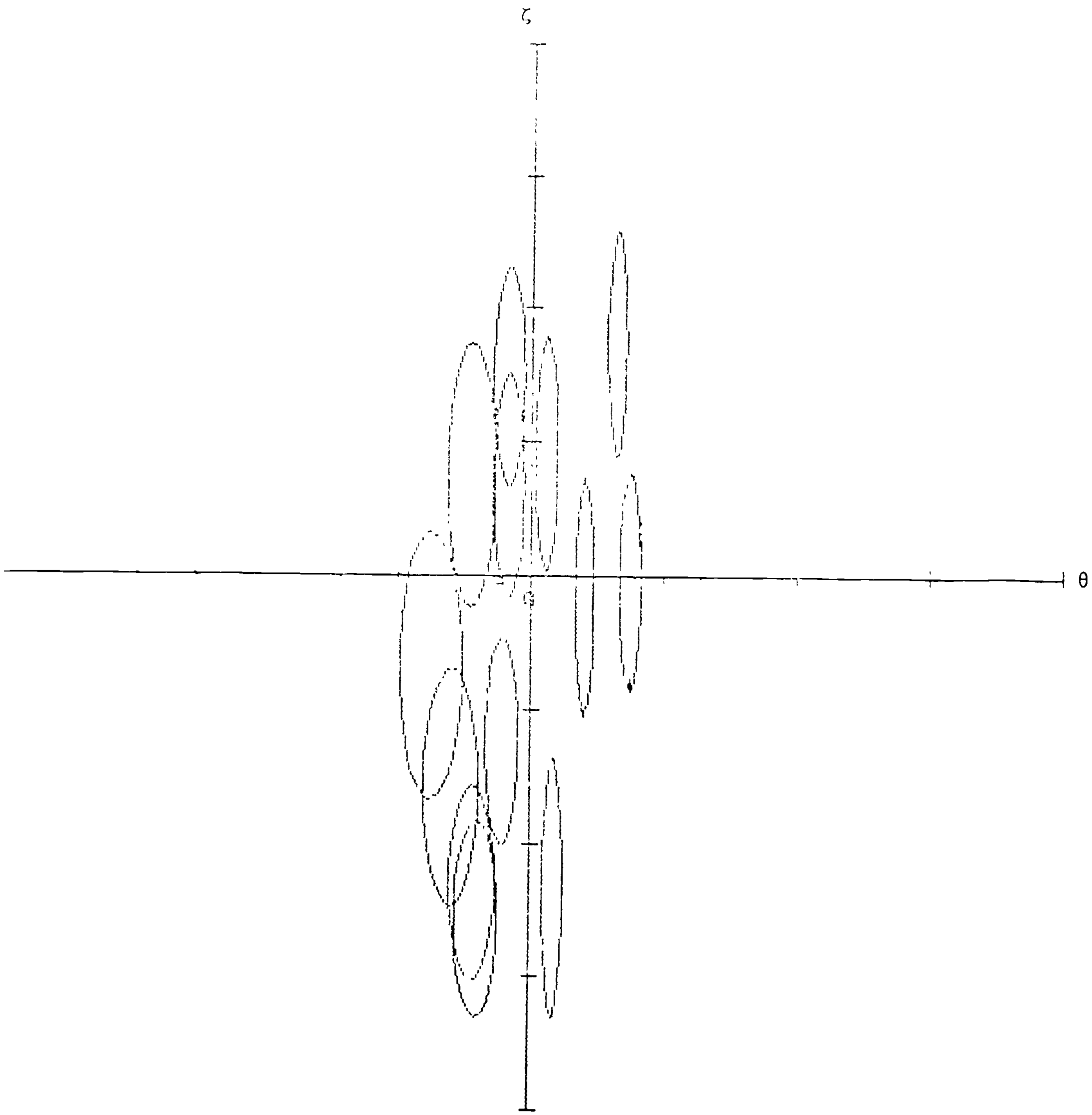

Figure 1: Fifteen Ellipses Representing Fifteen Error Types Randomly Chosen From Forty Sets of Ellipses