Appl. Statist. (2002)
51, *Part* 3, *pp.* 337–350

Data analytic methods for latent partially ordered classification models

Curtis Tatsuoka

George Washington University, Washington DC, USA

[Received January 2001. Final revision February 2002]

Summary. A general framework is presented for data analysis of latent finite partially ordered classification models. When the latent models are complex, data analytic validation of model fits and of the analysis of the statistical properties of the experiments is essential for obtaining reliable and accurate results. Empirical results are analysed from an application to cognitive modelling in educational testing. It is demonstrated that sequential analytic methods can dramatically reduce the amount of testing that is needed to make accurate classifications.

Keywords: Analysis of experiments; Cognitive modelling; Model fitting; Partially ordered set; Sequential classification

1. Introduction

Finite partially ordered classification models are useful for many statistical applications, including cognitive modelling. When the models are latent and complex, such as in cognitive applications, it becomes imperative to have available a variety of data analytic tools for fitting the models, and for the validation of assumptions that are made regarding the specification of class conditional response distributions for experiments. A data analytic framework for implementing latent finite partially ordered classification models is introduced which proposes such tools. This framework is illustrated with an example from educational testing in which sequential classification methods are applied to actual response data. It is demonstrated that sequential methods can drastically reduce the amount of testing that is needed to classify test subjects accurately to a true state within a finite partially ordered model.

Cognitive applications of interest include those in intelligent tutoring, educational testing and neuropsychological assessment. Finite partially ordered sets (posets) are natural models for cognition, as it is reasonable to assume that some cognitive states have higher levels of functionality than others. Poset models are flexible and can become quite rich and complex, enabling them to be effective models for describing response phenomena from test items or neuropsychological assessments. Importantly, they can provide concise and accurate information about cognitive functioning.

In the context of intelligent tutoring or educational testing, states in a poset model can be associated with information about cognitive attributes, such as whether or not they are mastered with respect to a given subject area (e.g. Tatsuoka (1995)). Such poset models may also be useful for neuropsychological assessment, when in clinical settings it is of interest to

Address for correspondence: Curtis Tatsuoka, Department of Statistics, George Washington University, Washington DC 20052, USA.
E-mail: tatsuoka@gwu.edu

 0035–9254/02/51337

determine the effect of a treatment or disease on cognitive functioning (e.g. Diabetes Control and Complications Research Group (1996) and Jaeger *et al*. (1992)).

Classification to a state is conducted by observing responses to experiments (e.g. test items). The relationship between responses and an underlying model is complex. For instance, in educational testing, a correct response on a test item gives evidence that the test subject has proficiency with respect to the cognitive attributes associated with the test item. An incorrect response by itself, however, does not necessarily indicate which of the cognitive attributes involved in the test item may be misunderstood. This situation is further complicated by random variation due to careless mistakes or lucky guesses. Also, since there may be multiple strategies for a test item in the sense that different sets of cognitive attributes can be relied on to perform well, it may be unclear which attributes a test subject has proficiency for even if their performance is good. Methodology for Bayesian sequential classification on poset models that can accurately diagnose test subjects in spite of this complexity has been presented in Tatsuoka and Ferguson (1999).

Motivation for this research stems from the work done in the field of educational testing by Tatsuoka and Tatsuoka (1987) and Tatsuoka (1990, 1995). Our current methodology differs from the approach described in the above references in that latent class conditional response distributions are based directly on a discrete poset model, and latent variable item response theory models (see Lord (1980)) are not used. Moreover, research on educational applications using poset models has been done by Falmagne, Doignon and others (e.g. Falmagne *et al*. (1990)). See also Macready and Dayton (1992), which describes sequential testing on a two-state poset (mastery *versus* non-mastery of a subject domain). The methods described below apply to abstract poset models, which leaves open the possibility for this methodology to be applied to other types of application, such as those in automated medical diagnosis or bioinformatics.

The data which are analysed in this paper can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

2. A case-study in educational testing

Let us now examine the implementation of latent poset classification models in educational testing. The test items and corresponding response data for this example were provided by Kikumi Tatsuoka and are referred to in Tatsuoka (1990). The subject domain involves the subtraction of fractions, and the analysis below is based on 20 test items that were submitted to 2144 students. The response distributions are Bernoulli. Both sequential and non-sequential classifications are conducted, where the objective is to identify a test subject's true state of cognition with respect to a given model. Two different poset models are fitted and then compared. Model I contains 59 states and model II contains 37 states, and the Hasse diagrams (see Davey and Priestly (1990)) of these models are shown in Figs 1 and 2 respectively. Each item is cognitively analysed in the sense that cognitive attributes are identified as being necessary for solving it correctly. In model I, eight attributes (or skills) are associated with the items (see Tatsuoka (1990)): 'Convert a whole number to a fraction', 'Separate a whole number from a fraction', 'Simplify before subtracting', 'Find a common denominator', 'Borrow from a whole number part', 'Column borrow to subtract the second numerator from the first', 'Subtract numerators' and 'Reduce the answer to its simplest form'. Denote these attributes respectively as A, B, ..., H. Model II is less complex than the first model and represents collections of the first seven of these attributes (attribute H is dropped). Items 10 and 12 include attribute H in their specifications relating to model I, but do not for model II. The attributes that are common to the two models will be denoted in the same manner. A partial list of what each of the knowledge states represents for model II

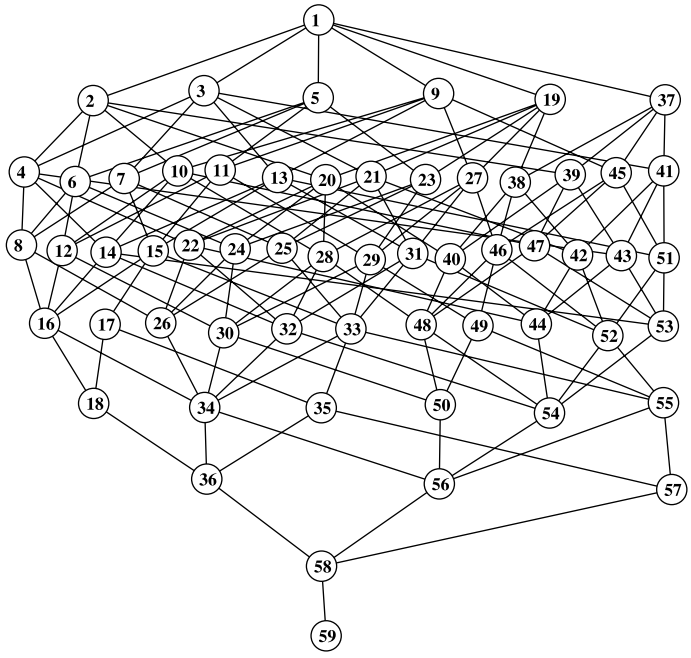

Fig. 1. Model I

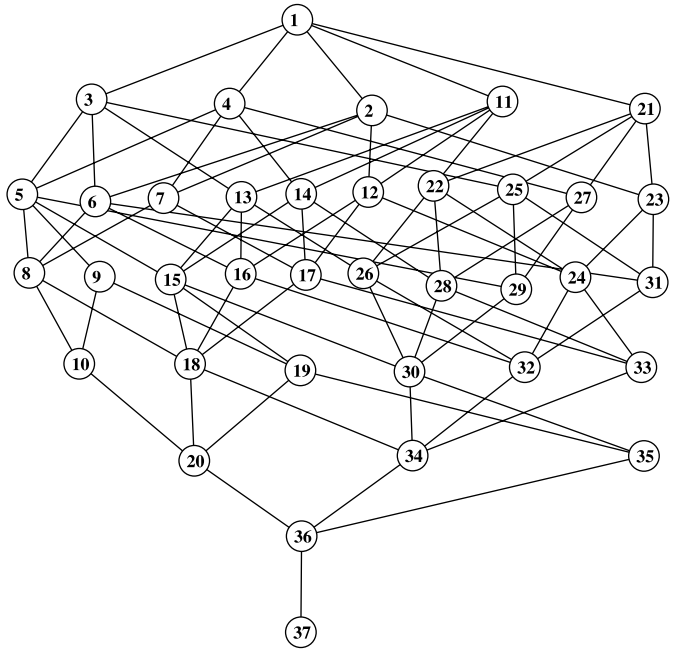

Fig. 2. Model II

Table 1. Partial list of knowledge states and cognitive information, model II

State	*Can*	*Cannot*
1	A B C D E F G	
2	B C D E F G	A
3	A B D E F G	C
4	A B C D F G	E
5	A B D F G	E and C
6	B D F G	C and A
7	B C D F G	C and A
8	B D F G	E and C and A

is given in Table 1. State 1 represents full knowledge of the domain (i.e. mastery of all seven attributes), whereas state 37 represents mastery of none of the skills. The other states represent partial knowledge, such as state 2, which represents 'can do attributes B, C, D, E, F and G, but not A'. After classification has been conducted, note that test subjects can be directed to focused remediation relating to specific attributes.

3. Statistical framework

Formally, the following statistical framework is adopted. Let S denote the set of classification states, one of which is the true state. Assume that S is a finite partially ordered set. Let $\mathcal{E}$ be a class of experiments, such as the pool of test items. A response random variable X given administration of experiment $e \in \mathcal{E}$ is assumed to have an associated class conditional response distribution $f(x|e, s)$ in relation to each $s \in S$. Suppose that $f(x|e, s)$ is a density with respect to a σ-finite measure μ_e on a measurable space $(\mathcal{X}_e, \mathcal{B}_e)$. Such a response could be a vector, such as when multiple-response variables are recorded from the same assessment. Once an experiment has been conducted, a response x is observed, and the information provided by the response is used in classifying the true state. A prior probability distribution for state membership within S, π_0, is assigned to each test subject. Let $\pi_0(j)$ represent the prior probability that a test subject is in $j \in S$. The posterior probability distribution on S after observing n responses is denoted by π_n. Conditionally on having chosen experiments $e'_1, e'_2, \ldots, e'_n \in \mathcal{E}$, and having respectively observed $X_1 = x_1,\ X_2 = x_2, \ldots,\ X_n = x_n$, the posterior probability that the test subject belongs to $j \in S$, $\pi_n(j)$, then becomes

$$\pi_n(j) \propto \pi_0(j) \prod_{i=1}^{n} f(x_i|e'_i, j). \tag{1}$$

A natural classification rule is to select the state in S with the largest terminal posterior probability value.

An experiment is said to *partition* a poset model in terms of subsets of states that share the same class conditional response distribution. A practical motivation for partitioning states to share response distributions is that it reduces the complexity of the response distribution estimation problem, which is particularly important when the underlying models are complex and latent. Each of the partitions can be interpreted as a subset of states that are equivalent in terms of how test subjects in those states will respond to a given experiment (e.g. test item). For cognitive applications, the specification of $f(x|e, s)$ can thus be a reflection of the cognitive

content of the experiment in relation to the underlying model. The up-set of an element y in a poset P is the set $\{z : y \leqslant z, z \in P\}$ (see Davey and Priestly (1990)). Let us denote this set by $\uparrow y$. For cognitive applications, a natural partition for an experiment could be an up-set and its complement. Given a state $y \in S$, the states in $\uparrow y$ are precisely the states that have at least as much associated functionality with respect to an experiment as y has. If X represents a response random variable for a given experiment $e \in \mathcal{E}$ with partition $\uparrow y$ and its complement, then the corresponding class conditional density for the true state $s \in S$ can be given by

$$f(x|e, s) = \begin{cases} f_e(x) & \text{if } y \leqslant s, \\ g_e(x) & \text{otherwise.} \end{cases}$$

As for multiple strategies, suppose that two states, y and z, represent different sets of cognitive attributes which can be used in different strategies to perform well on an experiment. A natural partition could then be the union of the up-sets that are generated by y and z.

For the fractions example, two partitions are adopted. Cognitively, the states are associated as either having the knowledge to answer an item correctly or not. Partitions consist either of an up-set or the union of up-sets, and the complement set. Let e_i correspond to item i in the item pool $\mathcal{E}$, $i = 1, \ldots, 20$. The response distribution for item $e_i \in \mathcal{E}$ in each partition is Bernoulli, with probability of success either $p_{\mathrm{u}}(e_i)$ or $p_{\mathrm{l}}(e_i)$. The value $1 - p_{\mathrm{u}}(e_i)$ can be interpreted as the probability of making a careless mistake, whereas $p_{\mathrm{l}}(e_i)$ can be viewed as the probability of making a lucky guess.

In practice, more complex partitions or response distributions can be used. For instance, in educational testing, multinomial response distributions can be employed for multiple-choice tests. It is possible that a particular erroneous response gives a strong indication that a student has a particular misconception or lacks a particular set of skills. Cognitive states identified as being especially prone to giving a particular erroneous response could comprise their own partition.

Two states are said to be *separated* by an experiment $e \in \mathcal{E}$ if the states are in different partitions. An order-preserving mapping of an initial poset can be defined such that its image is the collection of groups of states that cannot be separated by any $e \in \mathcal{E}$. It can be shown that this image also is a poset, with its order induced from the initial poset (see Tatsuoka (1996)). The models in Figs 1 and 2 are mappings from initial posets corresponding to the respective set of all possible subsets of attributes. Note that both model I and model II have fewer than 2^m states (the number of states in an initial poset), where $m = 8$ or $m = 7$ respectively. An implication is that information associated with states in these image posets may not provide clear information about all m attributes. For instance, for state 6 in model II, it is undetermined whether attribute E has been mastered or not.

4. Response distribution parameter estimation

Markov chain Monte Carlo estimation methods as described in Gelfand *et al.* (1992) were implemented to obtain item parameter estimates under order constraints $p_{\mathrm{u}}(e_i) > p_{\mathrm{l}}(e_i), e_i \in \mathcal{E}$. The responses are Bernoulli, and beta prior distributions $\beta\{\alpha_{\mathrm{u}}(e_i), \beta_{\mathrm{u}}(e_i)\}$ and $\beta\{\alpha_{\mathrm{l}}(e_i), \beta_{\mathrm{l}}(e_i)\}$ are adopted for the Bernoulli response parameters. Let $\mathbf{X}(k)$ be the response vector for test subject k. A Gibbs sampling procedure can be conducted as follows. The first step is to sample values for $p_{\mathrm{u}}(e_i)$ and $p_{\mathrm{l}}(e_i)$ from the respective prior distributions for the parameter values. Following Gelfand *et al.* (1992), parameter values are sampled on constrained regions with respect to the order relationship $p_{\mathrm{u}}(e_i) > p_{\mathrm{l}}(e_i)$. So, for example, on the basis of the most recently sampled value of $p_{\mathrm{u}}(e_i)$, $p_{\mathrm{l}}(e_i)$ is sampled from the constrained region $(0, p_{\mathrm{u}}(e_i))$. Let

$\theta = (p_{\mathrm{u}}(e_1), p_{\mathrm{l}}(e_1), p_{\mathrm{u}}(e_2), \ldots, p_{\mathrm{l}}(e_{20}))$ represent the most recently sampled parameter values, and let $N(k)$ represent the stopping time for test subject k (equal to 20 here). Given $\mathbf{X}(k)$ and θ, the second step is to calculate

$$\pi_{N(k)}|\mathbf{X}(k), \theta \tag{2}$$

where expression (2) represents the respective posterior distribution for state membership in S at stage $N(k)$ given $\mathbf{X}(k)$ and θ. Updating of the elements in expression (2) is conducted as in expression (1). The third step is then to sample from the corresponding distribution in expression (2) a value $s(k) \in S$ for each test subject. Let $\mathbf{X}_{\mathrm{u}}(e_i)$ and $\mathbf{X}_{\mathrm{l}}(e_i)$ correspond to the collection of responses to item e_i for test subjects with $s(k)$ in the respective partitions associated with $p_{\mathrm{u}}(e_i)$ and $p_{\mathrm{l}}(e_i)$. Fourth, for each $e_i \in \mathcal{E}$, we can then update prior distributions to obtain respective posterior beta distributions for the item parameters, $\beta\{p_{\mathrm{u}}(e_i)|\mathbf{X}_{\mathrm{u}}(e_i)\}$ and $\beta\{p_{\mathrm{l}}(e_i)|\mathbf{X}_{\mathrm{l}}(e_i)\}$. Item parameter values are then sampled again, and this process is repeated until convergence.

A total of 1930 of the responses from students were used in the estimation process, whereas the first 214 student responses in the data set were used to assess the classification performance on the basis of the resultant parameter estimates. This allocation of the data reflects the importance of obtaining accurate parameter estimates for the classification and data analytic processes. Although not done here, cross-validation methods could be employed. The beta distributions $\beta(1.5,1.0)$ and $\beta(1.0,1.5)$ were used respectively as priors for $p_{\mathrm{u}}(e)$ and $p_{\mathrm{l}}(e)$. These priors were chosen to be overdispersed over the range of respective expected possible values. Estimated posterior means for the parameters were used as the parameter estimates. The estimated item parameter values are given in Table 2 with corresponding specifications relating to model II. The estimates for model I are similar. The number of full cycle iterations within the Markov chain Monte Carlo chains was 100000 for both model I and model II. Convergence was deter-

Table 2. Item partitions and estimated item parameter values for model II†

Item	*Associated up-set*	*Estimated posterior mean*	*Estimated posterior standard deviation*	*Item*	*Associated up-set*	*Estimated posterior mean*	*Estimated posterior standard deviation*
1	{10}	0.13877	0.003266	11	{32}	0.06290	0.002347
		0.99979	0.006490			0.91036	0.007338
2	{20}	0.00021	0.000269	12	{36}	0.07455	0.009338
		0.97202	0.006577			0.85919	0.005726
3	{20}	0.00302	0.002146	13	{16}	0.01426	0.001903
		0.86582	0.005714			0.65379	0.005000
4	{24}	0.27413	0.002336	14	{34}	0.04475	0.011929
		0.88485	0.006034			0.95289	0.006749
5	{18,33}	0.22460	0.005007	15	{35}	0.07873	0.006144
		0.81186	0.005998			0.89891	0.007178
6	{36}	0.00056	0.000587	16	{34}	0.12406	0.007527
		0.95399	0.006736			0.89174	0.006754
7	{30}	0.02350	0.002940	17	{32}	0.03118	0.003663
		0.86588	0.009518			0.86598	0.006379
8	{36}	0.45855	0.006612	18	{29,31}	0.14449	0.002849
		0.80810	0.005405			0.99974	0.006491
9	{34}	0.32109	0.007002	19	{22}	0.02876	0.001594
		0.76194	0.005394			0.85959	0.006891
10	{32}	0.03266	0.007002	20	{24}	0.00656	0.001324
		0.75070	0.006311			0.92674	0.006551

†Per item, the first row corresponds to p_{l} and the second row to p_{u}.

mined by using the methods of Raftery and Lewis (1996) and also Geweke (1992). Estimated posterior means and standard deviations of the respective sampled values were calculated with the first 5000 iterations removed. An initial analysis based on a variety of prior distributions for the parameters and a smaller number of iterations yielded similar estimates. In Table 2, multiple associated up-sets for an item indicate that one of the partitions is the union of these up-sets.

5. Analysis of experiments

The data analytic tools that will be proposed involve analysing response distribution parameter estimates and patterns in the classification results. 'Global' measures for model assessment such as Bayes factors (e.g. Gelfand (1996)), although certainly useful, do not give direct insight into critical 'local' issues, such as the validation of partition specifications or the identification of specific aspects of a model where the fit may be poor.

The analysis of experiments (test items) entails identifying whether they efficiently and accurately provide information in classification. Desirable properties of an experiment include being able to discriminate clearly between states and having correctly specified partitions. For Bernoulli responses, discriminatory ability is gauged by how close the parameter of success values are to 1 and 0. The fastest possible (optimal) rates of convergence for $\pi_n(s) \to 1$ almost surely, $s \in S$ being the true state, are a function of the Kullback–Leibler information between response distributions (see Tatsuoka and Ferguson (1999)). This result implies that, the larger the information values, the faster the true state posterior probability value can converge to 1. These discrepancy values can be used as a diagnostic for validating partition specifications. Importantly, when the cognitive analysis (and hence partition specification) of a test item is incorrect, its estimated response distributions may not efficiently discriminate between separated states. For instance, suppose that a test item has two partitions, as in the case-study. If the Kullback–Leibler information is low, this could be because the partitions are contaminated in the sense that some states with high functionality could be included with the lower functionality states, or low functionality states could be included with the higher functionality states. Parameter estimates should indicate which of these situations is the case. In such a situation, the test item should be reanalysed cognitively, and partitions adjusted.

The following discussion specifically relates to the results for model II, since the estimated parameter values are similar for both models. For items 5 and 18, multiple strategies are identified, motivated in part to improve the estimation results. For instance, item 5 is of the form '$4\frac{3}{5} - 3\frac{4}{10} =$'. Two strategies that have been identified to solve this item are to use attributes A, B and G, or attributes B, C and G, and a partition of this item is adjusted to be the union of the up-sets of states 18 and 33. The results improved somewhat compared with specifying item 5 as just being associated with state 18 (i.e. one of the partitions is the up-set of state 18), as the p_{l}-values changed from 0.3123 to 0.2246, and the p_{u}-values changed from 0.7985 to 0.8119. Parameter estimates become slightly more attractive for item 18 as well. In Table 2, items 4, 8 and 9 were also found to have estimates that cause concern (see the respective p_{l}-estimates). Interestingly, item 8 is of the form '$\frac{2}{3} - \frac{2}{3} =$'. Apparently, students who did not have the skills specified by the respective test item partition still could identify that the correct answer was 0. It is well known that subtracting a number by itself leads to 0. Hence, alternative skills or knowledge that are not described by these models can probably be used to answer this item, and so this item does not discriminate between its specified partitions well. Like item 8, item 4 has an answer of 0. Similarly, item 9 is of the form '$3\frac{7}{8} - 2 =$'. This problem also is relatively easy in that it does not necessarily require the more complex attributes such as 'Converting a

whole number to a fraction (attribute A)' or 'Borrowing (attribute E)'. If it is of interest to test the given attributes, then these items are not as effective as others in the item pool. Item 13 has a low p_u-value, and it is of the form '$3\frac{3}{8} - 2\frac{5}{6}$'. It is one of the most difficult of the items, as it requires attributes B, D, E and G.

None-the-less, for both models, overall most of the estimated item values have excellent statistical properties, with corresponding p_u- and p_l-values being close to 1 and 0 respectively, and with estimates that appear to be stable. Statistically, these items discriminate between knowledge states in different partitions quite well. Also, it appears that in general the attributes identified describe the cognitive properties for the items adequately and that their partitions are correctly specified.

6. Model fitting diagnostics

Concerns in model fitting include determining whether a model has too many or not enough states. A model is too large when some of the states are superfluous and can be removed. A model is too small when some of the important underlying states are not included. An advantage to having a parsimonious model is that, for test subjects belonging to states that are specified, misclassification rates should be generally lower. Tatsuoka and Ferguson (1999) showed that, in addition to Kullback–Leibler information, optimal rates of convergence for the true state posterior probability to converge to 1 also depend on the complexity of the poset model around the true state. Specifically, complexity can be measured by the number of covers of a true state where, for y and z in a poset P, z is said to be a cover of y if $y < z$ and there does not exist a $w \in P$ such that $y < w < z$. Thus, for sequential classification, a parsimonious model does not require on average as many applications of experiments to stop with the same misclassification rate as a larger model that contains it. Moreover, since the optimal rates of convergence are exponential, they have an immediate effect on the classification error and on the length of sequentially administered testing sequences. Hence, it is imperative to use parsimonious models.

However, when models are too small, a dominant posterior probability may not emerge for test subjects belonging to missing states, and the classification error can be high. In sequential classification, this can lead to testing sequences that are relatively long, and decision theoretic loss values may be relatively higher. Hence, it is important that a model is sufficiently complex to represent the relevant states for a given application. Balancing parsimony while accurately representing the underlying classification model is a main task of model fitting.

An analysis of classification patterns is an important tool in assessing the fit of a model. For a true state $s \in S$, a necessary and sufficient condition for $\pi_n(s) \to 1$ almost surely is that s is separated from all $j \in S \backslash \{s\}$ infinitely often (see Tatsuoka and Ferguson (1999)). Using this result, classification phenomena can thus be predicted under various scenarios, such as when a valid state is not specified in the model (i.e. is *missing*), or when a state is superfluously included in a model. For instance, if a state is missing and a test subject belongs to that state, posterior mass probably will not aggregate on any one state. For each specified state in such a model, there will generally be some experiments in a given testing sequence for which the corresponding response distributions will differ from those of the true state if it were included. As the result indicates, responses to such experiments should lead to a decrease in the posterior probability for the state. The states that surround the location of the true state will generally be affected the least in this manner, so these states will share most of the posterior mass. It thus becomes possible to identify the location of missing states.

For both of the models in the example, the bottom portions of the models appear to be

sparse in terms of knowledge states, and there may be missing states. For model II, 25 students were classified non-sequentially to states 20, 34, 35 and 36. Six out of 25 were found to have a largest posterior probability value that was less than 0.44, and 11 out of the 25 were found to have a largest posterior probability value that was less than 0.70. These largest values are lower than those associated with the other states in the model as a whole, which may indicate that states are indeed missing from this portion of the model. The non-sequential classification results of these same 25 students in model I were overall quite similar.

If a state is superfluous, test subjects most probably will not be classified to that state. This follows since in general the experiments conducted will separate the true state from the other states in S, including the superfluous states, and hence responses should provide evidence in support of the true state. It thus becomes easy to identify potentially superfluous states, and they can be considered for removal from the model. For model II, with respect to non-sequential classification, all except three states are associated with the largest or second-largest posterior probability value for a student in the training sample. This indicates that model II does not contain many superfluous states. Model I had five states to which no students were classified or had a second-largest posterior probability value associated with it.

Finally, care should be given with respect to possibly aberrant response patterns. For instance, one student being classified answered the first nine test items correctly and then subsequently answered the remaining 11 items incorrectly. This could be an indication that the student stopped trying on the test. Such aberrant response patterns are candidates for removal from the data set (although that was not done here) and may not be due to problems with model fit or items.

7. Sequential classification

The sequential application of test items can be naturally implemented in the context of computer-based testing, when items can be administered one at a time. An important property of discrete poset classification models is that the optimal rates of convergence are exponential. This has profound ramifications in terms of the reduction in testing that can be achieved through sequential administration. The examples presented below illustrate the effectiveness of sequential methods in reducing the number of test items that are needed for an accurate classification. For model II, it generally takes only 4–9 test item responses for a posterior probability value associated with one state to satisfy the given stopping criterion of being greater than 0.80. As a comparison, note that latent variable item response theory models have $\sqrt{n}$ rates of convergence, as an underlying score value being estimated is assumed to lie within a continuous interval. It is likely that, at a minimum, 30 test items or so are required for accurate sequential estimation (e.g. Chang and Ying (1996)).

Consider now a decision theoretic formulation that can be adopted (see Ferguson (1967)). The parameter space of classification states and the set of terminal actions are both taken to be the finite poset S. Action $j \in S$ denotes that the test subject is classified into state j. For sequential classification, a cost of observation is assumed. The loss function can be written

$$L(s, j, n) = \begin{cases} 0 + cn & \text{if } j = s, \\ 1 + cn & \text{otherwise,} \end{cases}$$

where s is the true state in S, j is the action of choosing state j, n is the number of observations and $c > 0$ is the cost per observation. Of course, the cost of observation could vary depending

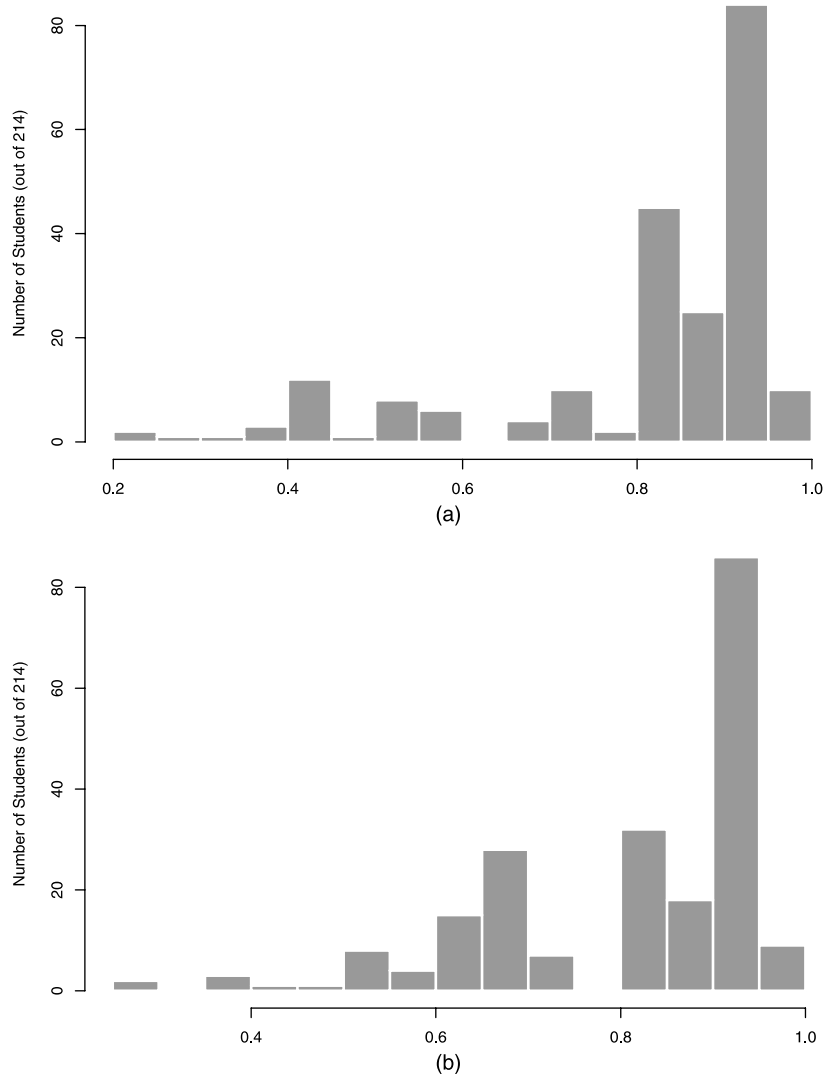

Fig. 3. Largest posterior probabilities for (a) sequential classification, model II, (b) sequential classification, model I, (c) non-sequential classification, model II, and (d) non-sequential classification, model I

on the experiment. The Bayes classification decision rule is to select the state with the largest terminal posterior probability value.

Experiments are selected for stage $n + 1$ on the basis of a rule that is a function of π_n. It is desirable for an experiment selection rule to attain the optimal rate of convergence almost surely. A class of experiment selection rules is shown to obtain the optimal rates of convergence almost surely under certain conditions in Tatsuoka and Ferguson (1999). The basis for one member of this class of experiment selection rules comes from information theory and was employed in the case-study. Denote Shannon entropy by

$$E\{n(\pi_n)\} = \sum_{j \in S} [-\log\{\pi_n(j)\}\, \pi_n(j)].$$

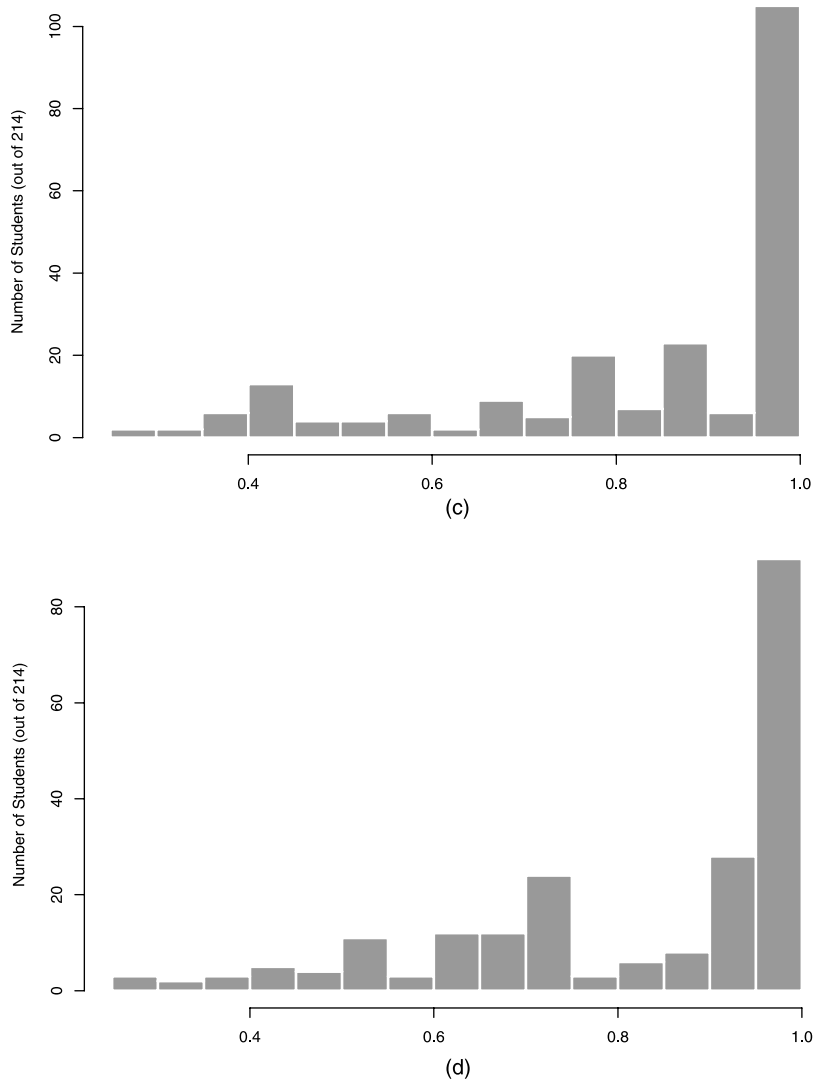

Fig. 3 (*continued*)

Define the *Shannon entropy procedure* sh: $\pi_n \to \mathcal{E}$ to select the experiment $e \in \mathcal{E}$ that minimizes

$$\mathrm{sh}(\pi_n, e) = \int E[n\{\pi_{n+1}(e, x)\}] \sum_{j \in S} \{f(x|e, j)\, \pi_n(j)\}\, \mu(\mathrm{d}x) \tag{3}$$

where $\pi_{n+1}(e, x)$ denotes the posterior probability distribution updated to stage $n + 1$ given $e_{n+1} = e$ and $X_{n+1} = x$, and μ is a σ-finite measure. This rule has the attractive feature of reducing the expected entropy among the posterior probability values for S and incorporating information about all the posterior probability values in π_n. However, it has been shown that the one-stage look ahead rule, which attempts to minimize the expected risk one stage ahead, may not perform well. This rule focuses only on the elements in π_n that subsequently have the potential to become the largest in π_{n+1} (see Tatsuoka and Ferguson (1999)).

Two stopping rules were used, with stopping being invoked if one of the rules called for it. One rule called for stopping when the largest posterior probability value exceeded 0.80, and the other was a two-stage look ahead stopping rule with cost of observation equal to 0.05. The latter stopping rule was employed only from the ninth stage of application, provided that it was reached. The stopping rules are somewhat aggressive, as is indicated by the short testing horizons.

In conducting the sequential classifications, test items were selected one stage at a time, and then the corresponding responses from the fixed set of responses were used as the observed response to the selected item. Figs 3(a) and 3(b) are histograms of the largest posterior probability values that are obtained through sequential testing for models II and I. The median largest posterior probability values are 0.8668 and 0.8681 respectively. As for the number of items being administered sequentially, the median value is 5 for model II and 7 for model I. The shortest testing horizon for model II is four items, whereas for model I it is five items. The longest sequences are 11 items and 10 items respectively. Overall, the size of the respective models is reflected in the lengths of the testing horizons, with sequential classification within the larger and more complex model generally requiring longer testing horizons. For model II, a sequence of an incorrect response to item 2 and subsequent correct responses to items 20, 18 and 7 leads to stopping and classification to state 21 with a corresponding posterior probability of 0.8147. Also, consecutive correct responses on items 2, 1, 7, 18 and 20 lead to classification to state 1 with a posterior probability equal to 0.9455, and consecutive incorrect responses to items 2, 20, 7, 14 and 6 lead to classification to state 37 with a posterior probability equal to 0.9017.

For the sequential classifications based on model II, the first quartile and median of the sums of the largest and second-largest posterior probability values are respectively 0.9253 and 0.9432. For model I, the comparable figures are 0.9418 and 0.9917 respectively. These figures indicate that, even if a classification is not decisive for one state, the posterior mass is mostly concentrated on at most two. This sum is given because a lack of clarification between two states can often be due to a lack of test items to separate them further. Still, the classification results can give precise information about the proficiency of particular attributes even if no clear dominant posterior probability emerges. For instance, attribute-specific probabilities of proficiency can be calculated by summing the posterior probabilities associated with states for which a particular attribute is mastered (see Tatsuoka (1995)).

For the non-sequential results, responses from all 20 items are used. Figs 3(c) and 3(d) give an indication of the magnitude of the terminal largest posterior probability values for the two models. For model II, there is a strong correspondence between sequential and non-sequential classification decisions, with 177 out of 214 matches between the classification decisions. For the remaining students, a natural measure of the discrepancy in the classification decisions is the cardinality of the symmetric difference between the subsets of attributes that are determined to be mastered. For the remaining 37 students who did not match, 29 of these non-matches had a cardinality of 1 with respect to the symmetric difference between the subsets of mastered attributes associated with the respective decisions. The mean cardinality of the symmetric differences among non-matches was 1.24. Because model II is smaller, stopping times in sequential classification are generally shorter, and this may be a cause for some of this lack of correspondence. In general, correspondence increases as the testing horizons are lengthened. For model I, 27 students did not have a match between their sequential and non-sequential decisions. Of those 27, 19 of these non-matches had a cardinality of 1 with respect to the symmetric difference. The mean cardinality among non-matches was 1.37. The strong correspondence between sequential and non-sequential classification decisions indicates that the specified stopping criteria are not too aggressive.

8. Summary of the case-study

The data analytic framework proposed has been applied to two models and used to compare them. This framework provides information about specific aspects of model fit such as the location of missing and superfluous states, and gives feed-back on the cognitive and statistical properties of each of the test items. Both models have similar parameter estimates and classification results. Model II is more parsimonious than model I, yet the fit for model II appears to be quite good in general. Moreover, for model II, owing to its smaller size, the performance of the sequential classification is generally more efficient in terms of the number of items that are administered until stopping. An analysis of the items gives an indication that the models describe the cognitive processes fairly well for most but not all of the items. It may be worthwhile to include more attributes in the models. However, a trade-off is that, as the underlying models become larger, at a certain point the loss values will probably increase for a large majority of students for which the additional attributes may not be needed to characterize them cognitively.

Importantly, multiple strategies are accounted for in the cognitive analysis and test item partition specifications. Also, data analysis indicates that complex underlying cognitive models are indeed needed to describe the predominant cognitive processes.

Acknowledgements

Special thanks are due to Dr Kikumi Tatsuoka, whose research forms the basis of this work. Thanks also go to Dr Ferenc Varadi for programming the routines, Dr Thomas Ferguson for all his help as my thesis advisor and Lin Yuan for computing assistance. Finally, thanks are due to the Associate Editor and the referees for helpful suggestions. This work was supported in part by National Science Foundation grant SES-9810202.

References

Chang, H.-H. and Ying, Z. (1996) A global information approach to computerized adaptive testing. *Appl. Psychol. Measmnt*, **20**, 213–229.

Davey, B. A. and Priestley, H. A. (1990) *Introduction to Lattices and Order*. Cambridge: Cambridge University Press.

Diabetes Control and Complications Research Group (1996) Effects of intensive diabetic therapy on neuropsychological function in adults in the diabetes control and complications trial. *Ann. Intern. Med.*, **124**, 379–388.

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P. and Johannesen, L. (1990) Introduction to knowledge spaces: how to build, test, and search them. *Psychol. Rev.*, **97**, 201–224.

Ferguson, T. (1967) *Mathematical Statistics: a Decision Theoretic Approach*. New York: Academic Press.

Gelfand, A. (1996) Model determination using sample-based methods. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.

Gelfand, A., Smith, A. and Lee, T.-M. (1992) Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Am. Statist. Ass.*, **87**, 523–532.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.

Jaeger, J., Berns, S., Tigner, A. and Douglas, E. (1992) Remediation of neuropsychological deficits in psychiatric populations: rationale and methodological considerations. *Psychpharm. Bull.*, **28**, 367–390.

Lord, F. (1980) *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale: Erlbaum.

Macready, G. B. and Dayton, M. C. (1992) The application of latent class models in adaptive testing. *Psychometrika*, **57**, 71–88.

Raftery, A. and Lewis, S. (1996) Implementing MCMC. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.

Tatsuoka, C. (1996) Sequential classification on partially ordered sets. *PhD Dissertation*. Cornell University, Ithaca.

Tatsuoka, C. and Ferguson, T. (1999) Sequential classification on partially ordered sets. *Technical Report 99-05*. Department of Statistics, George Washington University, Washington DC.

Tatsuoka, K. (1990) Toward an integration of item-response theory and cognitive error diagnosis. In *Diagnostic Monitoring of Skill and Knowledge Acquisition* (eds N. Frederiksen, R. Glaser, A. Lesgold and M. Shafto), pp. 453–488. Hillsdale: Erlbaum.

——— (1995) Architecture of knowledge structures and cognitive diagnosis: a statistical pattern classification approach. In *Cognitively Diagnostic Assessments* (eds P. Nichols, S. Chipman and R. Brennan), pp. 327–359. Hillsdale: Erlbaum.

Tatsuoka, K. and Tatsuoka, M. (1987) Bug distribution and statistical pattern classification. *Psychometrika*, **52**, 193–206.