PSYCHOMETRIKA—VOL. 36, NO. 4
DECEMBER, 1971

MATRIX ANALYSIS OF IDENTIFIABILITY OF SOME FINITE MARKOV MODELS*

JAMES G. GREENO

THE UNIVERSITY OF MICHIGAN

RICHARD B. MILLWARD

BROWN UNIVERSITY

COLEMAN T. MERRYMAN†

INDIANA UNIVERSITY

Methods developed by Bernbach [1966] and Millward [1969] permit increased generality in analyses of identifiability. Matrix equations are presented that solve part of the identifiability problem for a class of Markov models. Results of several earlier analyses are shown to involve special cases of the equations developed here. And it is shown that a general four-state chain has the same parameter space as an all-or-none model if and only if its representation with an observable absorbing state is lumpable into a Markov chain with three states.

The problem of identifiability refers to the possibility of uniquely determining a set of parameters of a model given complete information about the probabilities of all possible experimental outcomes. In any well-formulated model, there is a function λ that maps the parameter space of the model into the family of probability measures permitted by the model on the experimental outcome-space. That is, for each point in the parameter space, the probabilities of all possible experimental outcomes can be calculated. If some or all of the probability measures permitted by the model are associated with nonsingular sets of parameter points, then the model is not identifiable. In such a case, λ is not one-to-one and there is no inverse function λ^{-1} such that knowledge of all the probabilities of possible outcomes would specify a unique parameter. In an identifiable model, on the other hand, λ is one-to-one and λ^{-1} does exist. One common reason for a model to be nonidentifiable is that its parameter space is too large and consists of non-independent parameters. Another reason is that the outcome space, and thus the prob-

* This research was supported by the U.S. Public Health Service under Grant MH-12717 to Indiana University and Grant GM-1231 to the University of Michigan.

† Now at the University of Texas, Austin.

ability measure, is collapsed in such a way that distinctions between the parameter points are lost.

A number of recent papers have provided analyses of the identifiability of finite absorbing Markov chains that are useful in analyzing learning processes. Greeno and Steiner [1964, 1968] analyzed a general all-or-none learning model. Greeno [1967] analyzed a learning system involving short term retention, and showed that it is indistinguishable in data from simple all-or-none learning. Steiner and Greeno [1969] provided some general conditions under which models with several transient success states are indistinguishable in data from simple all-or-none learning. And Greeno [1968] analyzed the identifiability of a two-stage learning model. In each of these reports, the analysis has dealt with a specific model or a narrow class of models.

The purpose of the present paper is to present a general method that can be used to solve part of the problem of identifiability for any finite absorbing Markov chain that generates data in the form of sequences of correct responses and errors. The method is applied to the models analyzed in earlier papers, and the results clarify a principle that operates in all of the cases that have been studied.

A technique that has been useful in the earlier studies involves construction of a model with observable states. This concept was developed by Greeno and Steiner [1964, 1968], who defined an observable state as one whose occurrences can be specified on the basis of a data sequence. For example, if a model includes only one state in which errors occur, then that state is observable because each error in data corresponds to an occurrence of the error state in the model.

In studying the identifiability of a model $\mathfrak{M}$, it can be helpful to construct a model $\mathfrak{M}^*$ with observable states, such that the parameter space of $\mathfrak{M}^*$ generates all of the probability measures that are generated by the parameter space of $\mathfrak{M}$. When this happens, one says that $\mathfrak{M}$ implies $\mathfrak{M}^*$, since any data that satisfy $\mathfrak{M}$ also satisfy $\mathfrak{M}^*$.

The method that we present in this paper can be applied to any finite absorbing Markov model $\mathfrak{M}$ that generates binary data. The method permits construction of a model $\mathfrak{M}^*$ that is implied by $\mathfrak{M}$ and has an observable absorbing state. We proceed by first introducing notation for a general finite absorbing Markov model $\mathfrak{M}$ with initial probabilities Π and transition probabilities P, and a corresponding model $\mathfrak{M}^*$ with initial probabilities Π^* and transition probabilities P^*. The difference between the models is that in $\mathfrak{M}^*$ the absorbing state is observable while the absorbing state of $\mathfrak{M}$ is not. Then, using techniques developed by Bernbach [1966] and Millward [1969], we show that $\mathfrak{M}$ implies $\mathfrak{M}^*$, by developing a set of matrix equations that map the parameter space of $\mathfrak{M}$ into the parameter space of $\mathfrak{M}^*$.

Let μ designate the set of equations, and let Ω and Ω^* stand for the

parameter spaces of $\mathfrak{M}$ and $\mathfrak{M}^*$, respectively. Then

$$\mu : \Omega \rightarrow \Omega^*,$$

and μ is developed in such a way that for every parameter point ω in $\mathfrak{M}$'s parameter space (*i.e.*, $\omega \in \Omega$), $\mu(\omega) = \omega^* \in \Omega^*$ assigns the same probabilities to all events in data as does ω. In other words, μ is a function that gives values of new parameters but preserves probabilities of observable data.

The fact that μ exists shows that $\mathfrak{M}$ implies $\mathfrak{M}^*$, since every probability measure that is permitted by $\mathfrak{M}$ also is permitted by $\mathfrak{M}^*$. In applications, the functional relationship represented by μ can be used in analyzing the identifiability of $\mathfrak{M}$. The dimension of Ω^* is often lower than the dimension of Ω. If $\mathfrak{M}^*$ is identifiable, the dimension of Ω^* equals the number of identifiable parameter dimensions for $\mathfrak{M}$, and investigation of the inverse mapping of μ (that is, $\mu^{-1} : \Omega^* \rightarrow \Omega$) is very helpful in determining the kind of restrictions on $\mathfrak{M}$'s parameters that can be used to yield an identifiable version of $\mathfrak{M}$. In cases where $\mathfrak{M}^*$ is not identifiable, the dimension of Ω^* sets an upper limit on the number of identifiable dimensions for $\mathfrak{M}$, and investigation of the inverse mapping of μ can contribute to an understanding of certain aspects of the identifiability of $\mathfrak{M}$.

1. General Equations

We begin with a characterization of a general finite absorbing Markov model $\mathfrak{M}$, with an absorbing state D, where correct responses occur, an arbitrary number of transient error states $E_1, \cdots, E_e$, and an arbitrary number of transient success states $C_1, \cdots, C_c$. The initial and transition parameters of $\mathfrak{M}$ are denoted

$$\Pi = [D_{11}, E_{1e}, C_{1c}],$$

(1)

$$P = \begin{array}{c|c|c|c} & D & E_1 \cdots E_e & C_1 \cdots C_c \\ \hline D & 1 & 0 & 0 \\ \hline E_1 \\ \vdots \\ E_e & L_{e1} & F_{ee} & G_{ec} \\ \hline C_1 \\ \vdots \\ C_c & M_{c1} & H_{ce} & K_{cc} \end{array}$$

E_{1e} and C_{1c} are the $1 \times e$ and $1 \times c$ subvectors of initial probabilities for the transient error and correct states of $\mathfrak{M}$, respectively. F, G, H, K, L,

and M are submatrices of transition probabilities, with subscripts indicating the numbers of rows and columns in the various submatrices. $\mathfrak{M}$ is simply a general structure under which a wide variety of learning models fit as special cases.

We will investigate the relationship between $\mathfrak{M}$ and another general structure $\mathfrak{M}^*$. The difference between $\mathfrak{M}^*$ and $\mathfrak{M}$ is that the absorbing state of $\mathfrak{M}^*$ is observable, while $\mathfrak{M}$'s is not. The reason that State D in $\mathfrak{M}$ is not observable is that following the last error, there is an indefinite number of occurrences of the transient states $C_1, \cdots, C_c$, preceding entry into State D. To construct a model with an observable absorbing state we simply specify that the transitions from transient success states to the absorbing state are all zero. This new model $\mathfrak{M}^*$ has an absorbing state T, a set of transient error states $R_1, \cdots, R_e$, and a set of transient success states $S_1, \cdots, S_c$. The initial and transition probabilities of $\mathfrak{M}^*$ are denoted

$$\Pi^* = [T_{11}, R_{1e}, S_{1c}],$$

(2)

$$P^* = \begin{array}{c|c|c|c} & T & R_1 \cdots R_e & S_1 \cdots S_c \\ \hline T & 1 & 0 & 0 \\ \hline R_1 \\ \vdots \\ R_e & Z_{e1} & V_{ee} & W_{ec} \\ \hline S_1 \\ \vdots \\ S_c & 0 & X_{ce} & Y_{cc} \end{array}$$

To understand the relationship between the state spaces of $\mathfrak{M}$ and $\mathfrak{M}^*$, suppose (counter to fact) that it were possible to observe a sequence of the theoretical states of $\mathfrak{M}$, such as $E_1, E_3, C_2, C_1, E_2, C_3, C_2, D, D, \ldots$ This particular sequence would correspond to the following sequence of states of $\mathfrak{M}^*$: $R_1, R_3, S_2, S_1, R_2, T, T, T, T, \ldots$ In general, the outcome space for $\mathfrak{M}^*$ can be constructed from the outcome space for $\mathfrak{M}$ if, in each sequence that is in $\mathfrak{M}$'s outcome space we substitute R_i for each occurrence of E_i, substitute T for each correct response that follows the last error, and substitute S_i for each occurrence of C_i that precedes the last error. The correspondence between the outcome spaces amounts to a mapping from sets of outcomes in $\mathfrak{M}$'s outcome space into single outcomes in $\mathfrak{M}^*$'s outcome space. An important feature of the mapping is that within each equivalence class of $\mathfrak{M}$'s outcomes induced by the mapping, the different outcomes cannot be distinguished in data. Therefore, all of the observable

differences between outcomes in $\mathfrak{M}$'s outcome space are represented in $\mathfrak{M}^*$'s outcome space.

Of course, it is not the case that all of the states of $\mathfrak{M}^*$ are observable; specifically, the individual correct and error transient states are not observable except under special conditions. But the fact that T is observable turns out to make $\mathfrak{M}^*$ useful in analyzing the identifiability of several models.

We now proceed to develop expressions for μ, the function mapping Ω into Ω^*. As we mentioned earlier, μ consists of a set of matrix equations, and is constructed so as to preserve probability measures on data. For each submatrix and subvector of parameters of $\mathfrak{M}^*$, we develop an equation in terms of the submatrices and subvectors of the parameters of $\mathfrak{M}$. The equations are based on the correspondence between the outcome spaces of $\mathfrak{M}$ and $\mathfrak{M}^*$ described above. In general, probabilities of entering State T correspond to probabilities of entering State D, plus probabilities of entering the class of transient correct states and remaining there until absorption occurs. And probabilities of entering a transient correct state S_i correspond to probabilities of entering a transient correct state C_i and returning to one of the error states before absorption occurs. The reader may find it easier to sort out the notation in what follows by remembering that parameters in Ω are designated by letters in the first half of the alphabet (C through M, omitting I and J) and parameters in Ω^* are designated by letters in the last half of the alphabet (R through Z, omitting U).

First, consider the vector Z_{e1}. Each element in this vector is the probability of no more errors after an occurrence of a specified error state. From State E_i, the item may go directly to State D or it may go into the class of correct states, remaining there without an error until learning occurs. Thus, when $(I - K_{cc})^{-1}$ exists,

$$Z_{e1} = L_{e1} + G_{ec}(I - K_{cc})^{-1}M_{c1}. \tag{3}$$

Next, consider the submatrix V_{ee}. Since transitions among error states are not affected by the difference between the state spaces,

$$V_{ee} = F_{ee}. \tag{4}$$

To find W_{ec} we need to develop expressions for the individual elements. Let

$$\begin{aligned} w_{ij} &= P(S_{j,n+1} \mid R_{i,n}) \\ &= P(C_{j,n+1} \text{ and an error sometime after } n+1 \mid E_{i,n}), \end{aligned}$$

where the subscript n or $n + 1$ on the state indicates the trial number. Also let g_{ij} be the element of G_{ec} corresponding to w_{ij}. Finally, define the vectors H_i and K_i as the ith row vectors of the matrices H_{ee} and K_{cc}, respectively. Then

$$w_{ij} = g_{ij}H_i\Sigma_{e1} + g_{ij}K_i(I - K_{cc})^{-1}H_{ce}\Sigma_{e1} \tag{5}$$
$$= g_{ij}[H_i + K_i(I - K_{cc})^{-1}H_{ce}]\Sigma_{e1}$$

where Σ_{e1} is the $e \times 1$ unit vector for which postmultiplication is used to obtain the sum of terms in a $1 \times e$ vector. A complicated term in (5) arises later. We define

$$a_i = [H_i + K_i(I - K_{cc})^{-1}H_{ce}]\Sigma_{e1} ; \tag{6}$$

a_i is the probability of at least one error following an occurrence of state C_i. Then, instead of (5) we have

$$w_{ij} = g_{ij}a_j . \tag{7}$$

It is convenient to have an expression for W_{ee} in terms of matrices. Define $\Delta_{cc}(a_i)$ as the $c \times c$ diagonal matrix whose nonzero elements are a_1, a_2, $\cdots$, a_c. Then

$$W_{ee} = G_{ee}[\Delta_{cc}(a_j)]. \tag{8}$$

The remaining matrices in P^* are obtained as follows. Let

$$x_{ij} \,\varepsilon\, X_{ce} = P(E_{j,n+1} \mid C_{i,n} \text{ and an error sometime after } n)$$
$$y_{ij} \,\varepsilon\, Y_{cc} = P(C_{j,n+1} \mid C_{i,n} \text{ and an error sometime after } n).$$

Then

$$x_{ij} = \frac{h_{ij}}{a_i} \tag{9}$$
$$y_{ij} = \frac{k_{ij}a_j}{a_i}$$

where h_{ij} and k_{ij} are elements of H_{ce} and K_{cc}, respectively. The desired matrices can be obtained as

$$X_{ce} = [\Delta_{cc}(1/a_i)]H_{ce} , \tag{10}$$
$$Y_{cc} = [\Delta_{cc}(1/a_i)]K_{cc}[\Delta_{cc}(a_i)].$$

The initial probabilities can be obtained easily.

$$R_{1e} = E_{1e} . \tag{11}$$

To obtain S_{1c}, note that for $P(S_i) \,\varepsilon\, S_{1c}$ and $P(C_i) \,\varepsilon\, C_{1c}$,

$$P(S_i) = P(C_i)a_i .$$

Therefore,

$$(12) \qquad \begin{aligned} S_{1c} &= C_{1c}[\Delta_{cc}(a_i)], \\ T_{11} &= D_{11} + C_{1c}[\Delta_{cc}(1 - a_i)]\Sigma_{c1} . \end{aligned}$$

The derivation of μ is now complete. The function μ consists of equations (3), (4), (8), (10), (11), and (12). Any point $\omega \in \Omega$ generates a probability measure on the outcome space of $\mathfrak{M}$ which corresponds to a probability measure on an empirical outcome space of sequences of correct responses and errors. The point in Ω^* corresponding to ω, that is, $\omega^* = \mu(\omega)$, generates a probability measure on the outcome space of $\mathfrak{M}^*$ that also corresponds to a probability measure on the same empirical outcome space. And the two probability measures on the empirical outcome space are the same, since all of the observable differences between outcomes in $\mathfrak{M}$'s outcome space correspond to differences between outcomes in $\mathfrak{M}^*$'s outcome space.

The construction of $\mathfrak{M}^*$ according to the function μ represents a general solution for an important part of the problem of identifiability for finite absorbing Markov models. We have restricted outselves to models in which only correct responses occur after absorption. But within that class, any model with a finite number of transient states yielding only correct responses and errors can be translated into a theory with an observable absorbing state. As we will show below, when $e = c = 1$, construction of $\mathfrak{M}^*$ provides a direct solution of the identifiability problem. When either e or c is greater than one, construction of $\mathfrak{M}^*$ leaves some questions about identifiability to be dealt with, but it provides an important step in the analysis of identifiability. For example, (2) says something about the maximum number of identifiable parameters for $\mathfrak{M}$. The maximum number of parameters for $\mathfrak{M}$ is $(e + c)(e + c + 1)$, and the maximum number of parameters for $\mathfrak{M}^*$ is $(e + c)^2 + e$. (These quantities include the $e + c$ parameters of the initial vectors.) This does not imply that the number of identifiable parameters of any model will be c less than the number of theoretical parameters, but it does relate to the general fact that parameters may be nonidentifiable because of inability to distinguish in data between occurrences of an absorbing state and some occurrences of transient correct states. The specific implications of this general fact have to be worked out for individual models. The remainder of this paper consists of three relatively specific analyses, all of which use the methods developed above.

2. Simple All-or-None Learning

The results developed in Section 1 provide an immediate solution for the problem of identifiability when there is a single error state, and a single transient correct state. This is the model analyzed by Greeno and Steiner [1964] with transition probabilities

$$
(13) \qquad P = \begin{array}{c|ccc} & D & E_1 & C_1 \\ \hline D & 1 & 0 & 0 \\ E_1 & d & (1-d)t & (1-d)s \\ C_1 & c & (1-c)q & (1-c)p \end{array}
$$

where $t + s = p + q = 1$.

Using (3) and (4),

$$
(14) \qquad \begin{aligned} Z_{11} &= (d + (1-d)s[1 - (1-c)p]^{-1}c) \\ &= \underline{\left(d + \frac{(1-d)sc}{q+pc}\right)} \\ V_{11} &= ((1-d)t). \end{aligned}
$$

From (6)

$$
\begin{aligned} a_1 &= (1-c)q + (1-c)p[1-(1-c)p]^{-1}(1-c)q \\ &= \frac{(1-c)q}{q+pc}. \end{aligned}
$$

Then, with (7)–(10),

$$
(15) \qquad \begin{aligned} W_{11} &= (w_{11}) = \left(\frac{(1-d)s(1-c)q}{q+pc}\right) \\ X_{11} &= (x_{11}) = (q+pc), \\ Y_{11} &= (y_{11}) = ((1-c)p). \end{aligned}
$$

The initial probabilities are obtained using (11) and (12).

$$
(16) \qquad \begin{aligned} R_{11} &= E_{11}\,, \\ S_{11} &= \frac{C_{11}(1-c)q}{q+pc} \\ T_{11} &= D_{11} + \underline{C_{11}\left[1 - \frac{(1-c)q}{q+pc}\right]} \\ &= D_{11} + \frac{C_{11}c}{q+pc}. \end{aligned}
$$

Thus, in this case $\mathfrak{M}^*$ has the initial probabilities given in (16), and transition probabilities

$$
(17) \qquad P^* = \begin{array}{c|ccc} & T & R & S \\ \hline T & 1 & 0 & 0 \\ R & d + \dfrac{(1-d)sc}{q+pc} & (1-d)t & \dfrac{(1-d)s(1-c)q}{q+pc} \\ S & 0 & q+pc & (1-c)p \end{array}
$$

in agreement with Greeno and Steiner [1964, equation (43)].

Equations (16) and (17) show that the model of (13) can have at most five identifiable parameters; two in the initial vector and three in the transition matrix. These may be labelled

$$
\begin{aligned}
\pi &= T_{11} = D_{11} + \frac{C_{11}c}{q+pc} \\
\theta &= R_{11}/(1 - T_{11}) = \frac{E_{11}}{E_{11} + C_{11}\left(\dfrac{q(1-c)}{q+pc}\right)}, \\
(18) \qquad u &= Z_{11} = \left(d + \frac{(1-d)sc}{q+pc}\right), \\
v &= V_{11}/(1 - Z_{11}) = \left(\frac{t(q+pc)}{q + c(p-s)}\right), \\
w &= X_{11} = (q + pc).
\end{aligned}
$$

Now that the identifiable parameters are known, it is possible to consider restrictions on and dependencies among the original theoretical parameters. If Ω is larger than Ω^*, then some of the theoretical parameters must be related. One important result of this analysis is a clear specification of such dependencies. Restrictions on the theoretical parameters lead to two possible results: (1) Testable Assumptions. The restrictions on the parameters in Ω can, in turn, yield a relationship between the identifiable parameters in Ω^*, and this relationship is open to experimental tests. (2) The restrictions can impose no restrictions on the identifiable parameters. In this case they simply serve to select a particular model or class of models. If a restriction on the theoretical parameters has the property that the theoretical parameters are identifiable when the restriction is imposed, the restriction is called an identifiable restriction. Greeno and Steiner showed that if $p = s$ is assumed in the theoretical parameters of (13), then the identifiable parameters must satisfy $v = w$, hence, $p = s$ is a testable assumption. However, restrictions involving the learning parameters such as $c = d$ or $c = 0$ do not generally impose restrictions on the identifiable parameters; hence, these are identifying restrictions.

Greeno and Steiner presented very little analysis of assumptions about the initial vector. A few brief comments can be presented here to make that picture more complete. We will examine three assumptions about the initial probabilities.

First, it often is natural to assume that no items begin in State D. That is, $D_{11} = 0$. If this assumption is made, we have

$$\pi = \frac{C_{11}c}{q + pc}$$

$$\theta = \frac{1 - C_{11}}{1 - \frac{C_{11}c}{q + pc}} \tag{19}$$

Since $q + pc = w$, an identifiable parameter, (19) shows that assuming $D_{11} = 0$ would permit estimates of C_{11} and c. Hence, $D_{11} = 0$ is an identifying restriction, not a testable assumption.

Next, consider the possibility that some items may start in State D, but of those that do not the probability of a correct response on the first trial is g, a known guessing probability. This imposes the restriction

$$\frac{C_{11}}{C_{11} + E_{11}} = g, \quad \text{or} \quad C_{11} = \frac{g}{1 - g} E_{11},$$

which leads to the relationship

$$\theta = \frac{1 - g}{1 - \frac{gc}{q + pc}}. \tag{20}$$

Equation (20) shows that the assumption of guessing on the first trial permits us to estimate c. Hence, this too is an identifying restriction, not a testable assumption.

Finally, consider the possibility that for items not in State D at the beginning, the probability of a correct response is p, the probability of a correct response on later trials following correct responses. This imposes the restriction

$$C_{11} = \frac{p}{q} E_{11},$$

which leads to the relationship

$$\theta = w. \tag{21}$$

Thus, this restriction is a testable assumption, and if it is accepted the number of identifiable parameters is reduced.

3. *All-or-None Learning with Short Term Retention*

A somewhat more complicated case was analyzed by Greeno [1967]. The model is given in (22); E_1 and C_1 represent states producing wrong and correct responses, respectively, and C_2 represents short-term retention of an item. Again, D is the learned state.

$$\Pi = [t,\ (1-t)(1-b)(1-g),\ (1-t)(1-b)g,\ (1-t)b]$$

(22)

$P =$	D	E_1	C_1	C_2
D	1	0	0	0
E_1	$d+(1-d)a$	$(1-d)(1-a)(1-h)(1-g)$	$(1-d)(1-a)(1-h)g$	$(1-d)(1-a)h$
C_1	$d+(1-d)a$	$(1-d)(1-a)(1-h)(1-g)$	$(1-d)(1-a)(1-h)g$	$(1-d)(1-a)h$
C_2	$c+(1-c)a$	$(1-c)(1-a)(1-h)(1-g)$	$(1-c)(1-a)(1-h)g$	$(1-c)(1-a)h$

To make calculations simpler, we use the notation of (1).

(23)

$P =$	D	E_1	C_1	C_2
D	1	0	0	0
E_1	ℓ_{11}	f_{11}	g_{11}	g_{12}
C_1	m_{11}	h_{11}	k_{11}	k_{12}
C_2	m_{21}	h_{21}	k_{21}	k_{22}

We will illustrate the algebraic work by sketching the application of (3). We have

$$(I - K_{22})^{-1} = \frac{1}{(1 - k_{11})(1 - k_{22}) - k_{12}k_{21}} \begin{bmatrix} 1 - k_{22} & k_{12} \\ k_{21} & 1 - k_{11} \end{bmatrix}$$

Then

$$Z_{11} = \left[\ell_{11} + \frac{(g_{11}g_{12}) \begin{bmatrix} 1 - k_{22} & k_{12} \\ k_{21} & 1 - k_{11} \end{bmatrix} \begin{bmatrix} m_{11} \\ m_{21} \end{bmatrix}}{(1 - k_{11})(1 - k_{22}) - k_{12}k_{21}}\right]$$

This gives

$$Z_{11} = \left(1 - \left[\frac{(1 - d)(1 - a)(1 - h)(1 - g)}{1 - h(1 - c)(1 - a) - (1 - h)g(1 - d)(1 - a)}\right]\right) \tag{24}$$

in agreement with Greeno [1967, equation (7)].

Carrying out the remaining algebraic work, we obtain the following:

$$V_{11} = ((1 - d)(1 - a)(1 - h)(1 - g)). \tag{25}$$

The elements of W_{12} are

$$w_{11} = \frac{(1-d)^2(1-a)^2(1-h)^2g(1-g)}{1-h(1-c)(1-a)-(1-h)g(1-d)(1-a)}$$
$$w_{12} = \frac{(1-d)(1-a)^2h(1-c)(1-h)(1-g)}{1-h(1-c)(1-a)-(1-h)g(1-d)(1-a)} \tag{26}$$

The elements of X_{21} are equal.

$$x_{11} = x_{21} = 1 - h(1-c)(1-a) - (1-h)g(1-d)(1-a). \tag{27}$$

And the four elements of Y_{22} are

$$y_{11} = y_{21} = (1-d)(1-a)(1-h)g$$
$$y_{12} = y_{22} = (1-c)(1-a)h. \tag{28}$$

The equalities in the X_{21} and Y_{22} matrices provide an explanation for another part of the identifiability problem in this model. To see this, consider the following representation of (2).

$$P^* = \begin{array}{c|cccc} & T & R_1 & S_1 & S_2 \\ \hline T & 1 & 0 & 0 & 0 \\ R_1 & z_{11} & v_{11} & w_{11} & w_{12} \\ S_1 & 0 & x_{11} & y_{11} & y_{12} \\ S_2 & 0 & x_{21} & y_{21} & y_{22} \end{array} \tag{29}$$

Equation (28) shows that the two bottom rows are identical; thus, states S_1 and S_2 can be combined into a single state, and the resulting process will still be a Markov chain [Burke & Rosenblatt, 1958]. And since the response that occurs in S_1 is the same as in S_2, we will not be able to distinguish the lumped Markov chain from the one described by (29). That is, the model may be expressed as

$$P^* = \begin{array}{c|ccc} & T & R_1 & S_\cdot \\ \hline T & 1 & 0 & 0 \\ R_1 & z_{11} & v_{11} & w_{11} + w_{12} \\ S_\cdot & 0 & x_{\cdot 1} & y_{\cdot 1} + y_{\cdot 2} \end{array} \tag{30}$$

without losing any of its empirical content. But there is one additional constraint. From (24), (25), and (27), it may be noted that

$$(1 - z_{11})x_{\cdot 1} = v_{11}\,, \tag{31}$$

which means that only two parameters are needed to specify the empirical transition probabilities. Equation (31) corresponds to having $v = w$ in the notation of (18). Two more parameters are needed to specify the initial vector, so the model has four identifiable parameters. Greeno [1967] pointed out that $b = 0$ and $b = h$ are testable assumptions, and that hypotheses about t are testable under either of these assumptions about b.

It turns out, then, that (22) has the same identifiable parameter space as does a model of simple all-or-none learning. The identifiable parameters of (22) are the same as those of (13) when $p = s$ in the latter. It was not obvious at the outset that this would be the case. In (22), States C_1 and C_2 are not lumpable. However, the present analysis, using the matrix equations of Section 1, makes it clear why this addition of a short term memory state does not add to the number of identifiable parameters in this model. Once the model is put into a form with an observable absorbing state, as in (29), then the two transient states leading to correct responses are lumpable.

A more general analysis of models of this type was given by Steiner and Greeno [1969]. The model analyzed was

(32)
$$P = \begin{array}{c|cccc} & D & E & C_1 & C_2 \\ \hline D & 1 & 0 & 0 & 0 \\ E & P_{03} & P_{00} & P_{01} & P_{02} \\ C_1 & P_{13} & P_{10} & P_{11} & P_{12} \\ C_2 & P_{23} & P_{20} & P_{21} & P_{22} \end{array}$$

Steiner and Greeno showed that (32) has the same identifiable parameter space as an all-or-none model if and only if there exists a β between zero and one such that

(33)
$$[K_{22} - \Delta_{22}(\beta)]H_{21} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

where K_{22} and H_{21} are the matrices defined in (1) and $\Delta_{22}(\beta)$ is the diagonal matrix with the diagonal elements both equal to β. Equation (33) is satisfied if H_{21} is an eigenvector of K_{22} corresponding to the eigenvalue β.

> (Steiner and Greeno's analysis was restricted to responses starting with the first error. This effectively suppresses considerations involving initial probabilities. Analysis taking all possible initial vectors into account is usually cumbersome, and often can be carried out more usefully in connection with specific applications.)

When the entries in (32) are substituted in (33), we obtain the following equations:

$$P_{21}P_{10} + P_{22}P_{20} = P_{20}\beta, \qquad P_{11}P_{10} + P_{12}P_{20} = P_{10}\beta. \tag{34}$$

The most interesting case has P_{10} and P_{20} nonzero. In that case, (33) is satisfied if and only if

$$P_{11} + P_{12}\left(\frac{P_{20}}{P_{10}}\right) = P_{22} + P_{21}\left(\frac{P_{10}}{P_{20}}\right), \tag{35}$$

and when (35) is satisfied the value of each expression equals β.

Recall that (33) must be satisfied with $\beta \leq 1$. We can show that if (35) is satisfied, then $\beta \leq 1$. Suppose that the left side is greater than one. Then

$$\left(\frac{P_{20}}{P_{10}}\right) > \left(\frac{1 - P_{11}}{P_{12}}\right) = \left(\frac{P_{12} + P_{13} + P_{10}}{P_{12}}\right) > 1,$$

since we have assumed that P_{10} is nonzero. Therefore,

$$\left(\frac{P_{10}}{P_{20}}\right) < 1. \tag{36}$$

Now consider the right side of (35).

$$P_{22} + P_{21}\left(\frac{P_{10}}{P_{20}}\right) < 1 - P_{21} + P_{21}\left(\frac{P_{10}}{P_{20}}\right) < 1.$$

because of (36). Thus, if the left side of (35) is greater than one, the right side is less than one; hence, (35) cannot be satisfied unless $\beta \leq 1$.

Additional cases satisfying (34) and the resulting values of β are

$$\begin{aligned} &P_{10} = P_{12} = 0, \quad P_{20} > 0; \quad \beta = P_{22}\,. \\ &P_{20} = P_{21} = 0, \quad P_{10} > 0; \quad \beta = P_{11}\,. \\ &P_{10} = P_{20} = 0; \quad \beta \text{ arbitrary}. \end{aligned} \tag{37}$$

Equations (35) and (37) give the conditions under which (32) has the same parameter space as an all-or-none model. Earlier, we showed that (22) has the same parameter space as an all-or-none model. The earlier argument was based on (27) and (28), which showed that (29) was lumpable to a theory with just one transient success state.

Equations (27) and (28) apply only to a specific model. We now show a more general result. Let $\mathfrak{M}$ be a model in the form of (32), and let $\mathfrak{M}^*$ be the model obtained by applying the function μ, given in Section 1. Note that $\mathfrak{M}^*$ will have the form of (29). If P_{10} and P_{20} are both nonzero in $\mathfrak{M}$, then $\mathfrak{M}$ has the same identifiable parameter space as an all-or-none model if and only if S_1 and S_2 are lumpable in $\mathfrak{M}^*$.

Equation (35) gives the condition for $\mathfrak{M}$ to have the same parameter space as an all-or-none model, when x_{11} and x_{21} are equal in (29). Thus, to

prove the claim made above, we need to show that (35) is satisfied if and only if $x_{11} = x_{21}$ in (29).

First, we calculate the value of x_{11} . The relevant equations are (6) and (9), where

$$H_1 = P_{10}\,,\ K_1 = (P_{11}P_{12}),\ H_{21} = \begin{bmatrix} P_{10} \\ P_{20} \end{bmatrix},$$

$$(I - K_{22})^{-1} = \begin{bmatrix} 1 - P_{11} & -P_{12} \\ -P_{21} & 1 - P_{22} \end{bmatrix}^{-1}$$

$$= \frac{1}{(1 - P_{11})(1 - P_{22}) - P_{12}P_{21}} \begin{bmatrix} 1 - P_{22} & P_{12} \\ P_{21} & 1 - P_{11} \end{bmatrix}$$

$(I - K_{22})^{-1}$ will always exist in the case we are considering. The condition we need is

$$(1 - P_{11})(1 - P_{22}) - P_{12}P_{21} \neq 0.$$

$$(1 - P_{11})(1 - P_{22}) - P_{12}P_{21} = (P_{10} + P_{12} + P_{13})(P_{20} + P_{21} + P_{23}) - P_{12}P_{21}$$

which is clearly greater than zero, since P_{10} and P_{20} are nonzero. Now, apply (6),

$$a_1 = \frac{P_{10}(1 - P_{22}) + P_{12}P_{20}}{(1 - P_{11})(1 - P_{22}) - P_{12}P_{21}}.$$

Then, by (9),

$$x_{11} = \frac{P_{10}}{a_1} = \frac{P_{10}[(1 - P_{11})(1 - P_{22}) - P_{12}P_{21}]}{P_{10}(1 - P_{22}) + P_{12}P_{20}}. \tag{38}$$

Similar calculations lead to

$$x_{21} = \frac{P_{20}[(1 - P_{11})(1 - P_{22}) - P_{12}P_{21}]}{P_{20}(1 - P_{11}) + P_{21}P_{10}}. \tag{39}$$

Now, if (35) is satisfied, simple algebra yields

$$P_{12} = \left(\frac{P_{10}}{P_{20}}\right)\left[P_{22} - P_{11} + P_{21}\left(\frac{P_{10}}{P_{20}}\right)\right].$$

Substituting this in the denominator of (38),

$$x_{11} = \frac{(1 - P_{11})(1 - P_{22}) - P_{12}P_{21}}{1 - P_{11} + P_{21}\left(\frac{P_{10}}{P_{20}}\right)}$$

Thus, if (32) has the same identifiable parameters as an all-or-none model, then S_1 and S_2 are lumpable in (29).

Working the other way, since $(1 - P_{11})(1 - P_{22}) - P_{12}P_{21}$ is nonzero, if $x_{11} = x_{21}$,

$$\frac{P_{10}}{P_{10}(1 - P_{22}) + P_{12}P_{20}} = \frac{P_{20}}{P_{20}(1 - P_{11}) + P_{21}P_{10}}.$$

Equation (35) is then obtained by taking reciprocals on both sides and carrying out simple algebra. Thus, if S_1 and S_2 are lumpable in (29), then (32) has the same identifiable parameters as an all-or-none model.

Recall that the above argument applies when P_{10} and P_{20} are nonzero. When any of the conditions of (37) occur, different kinds of lumpability arise. For example, when

$$P_{10} = P_{12} = 0, \qquad P_{20} > 0,$$

(32) becomes

(40)
$$P = \begin{array}{c|cccc} & D & E & C_1 & C_2 \\ \hline D & 1 & 0 & 0 & 0 \\ E & P_{03} & P_{00} & P_{01} & P_{02} \\ C_1 & P_{13} & 0 & P_{11} & 0 \\ C_2 & P_{23} & P_{20} & P_{21} & P_{22} \end{array}$$

States D and C_1 constitute an absorbing class, and can be lumped to form the chain

(41)
$$P' = \begin{array}{c|ccc} & D' & E & C_2 \\ \hline D' & 1 & 0 & 0 \\ E & P_{01} + P_{03} & P_{00} & P_{02} \\ C_2 & P_{21} + P_{23} & P_{20} & P_{22} \end{array}$$

When the function μ is applied to (41), the result clearly will be a chain with only three states. Similar degeneracies occur in the other cases given in (37). Thus, we conclude that the following theorem holds:

A model $\mathfrak{M}$ of the form of equation (32) has the same parameter space as an all-or-none model, if and only if application of μ yields a model $\mathfrak{M}^*$ which is lumpable into a Markov chain with three states.

An interesting special case of the all-or-none parameters often occurs. This is the case where $v = w$, using the notation of (18). Equation (22) satisfies this condition, as we remarked earlier. If the more general model given as (32) reduces to an all-or-none model, then it will satisfy the further

restriction of $v = w$ if equation (31) is satisfied. Carrying out the calculations using (3) and (5), this happens if

$$1 = x_{.1} + \frac{P_{01}P_{10} + P_{02}P_{20}}{P_{00}}$$

where $x_{.1}$ has the value given in (38) and (39).

4. Two-Stage Learning with No Successes in the Initial Stage

The final model to be discussed here was analyzed by Greeno [1968]. It is given in (42); E_2 is the initial state, E_1 and C_1 represent errors and correct responses in the intermediate state, and D is the learned state.

$$\Pi = [t, (1 - s - t)r, s, (1 - s - t)(1 - r)],$$

(42)

$$P = \begin{array}{c|cccc} & D & E_1 & E_2 & C_1 \\ \hline D & 1 & 0 & 0 & 0 \\ E_1 & d & (1-d)q & 0 & (1-d)p \\ E_2 & ab & a(1-b)e & 1-a & a(1-b)(1-e) \\ C_1 & c & (1-c)q & 0 & (1-c)p \end{array}$$

When we apply the function μ we obtain values for parameters of $\mathfrak{M}^*$ as follows:

(43)

$$T_{11} = \left(t + \frac{(1 - s - t1 - r)c}{q + pc}\right),$$

$$R_{21} = \begin{bmatrix} (1 - s - t)r \\ s \end{bmatrix}.$$

$$S_{11} = \left(\frac{(1 - s - t)(1 - r)(1 - c)q}{q + pc}\right),$$

$$Z_{21} = \begin{bmatrix} \dfrac{qd + pc}{q + pc} \\ ab + \dfrac{a(1 - b)(1 - e)c}{q + pc} \end{bmatrix}$$

$$V_{22} = \begin{bmatrix} (1 - d)q & 0 \\ a(1 - b)e & 1 - a \end{bmatrix},$$

$$W_{21} = \begin{bmatrix} \dfrac{(1 - d)p(1 - c)q}{q + pc} \\ \dfrac{a(1 - b)(1 - e)(1 - c)q}{q + pc} \end{bmatrix},$$

$$X_{12} = (q + pc \quad 0),$$

$$Y_{11} = ((1 - c)p).$$

When (43) is arranged as an initial vector and a transition matrix, it has the form

$$\Pi = (t_1 \, , r_1 \, , r_2 \, , s_1)$$

(44)

$$P^* = \begin{array}{c|cccc} & T & R_1 & R_2 & S_1 \\ \hline T & 1 & 0 & 0 & 0 \\ R_1 & z_{11} & v_{11} & 0 & w_{11} \\ R_2 & z_{21} & v_{21} & v_{22} & w_{21} \\ S_1 & 0 & x_{11} & 0 & y_{11} \end{array}$$

There are nine theoretical parameters in (42). In (44) it appears that there might be nine identifiable parameters. However, Greeno [1968] showed that there are only seven identifiable parameters. The reduction occurs because of dependencies among the various quantities in (43).

The first of these is straightforward.

$$x_{11} = \frac{v_{11}}{1 - z_{11}};$$

that is

$$q + pc = \frac{(1 - d)q}{1 - \dfrac{qd + pc}{q + pc}}.$$

The remaining restriction is more devious. To explain how it comes about, we consider a system that has two transient states and an absorbing state,

$$P_1(\underline{A}, R_1 \, , R_2) = (0, \, 1 - \delta, \, \delta),$$

(45)

$$P = \begin{array}{c|ccc} & A & R_1 & R_2 \\ \hline A & 1 & 0 & 0 \\ R_1 & \gamma & 1 - \gamma & 0 \\ R_2 & \alpha\beta & \alpha(1 - \beta) & 1 - \alpha \end{array}$$

If R_1 and R_2 represent indistinguishable error states, and A is entered on the trial of the first correct response, then (45) will produce data corresponding to the initial string of errors produced by (42).

Let X be the number of trials before absorption; that is, State A is entered on Trial $X + 1$. The probability distribution of X constitutes the only observable data that (45) will produce. When $\gamma \neq \alpha$, the probability distribution is

$$(46) \quad P(X = n) = \delta\left(\frac{\gamma - \beta\alpha}{\gamma - \alpha}\right)\alpha(1 - \alpha)^{n-1} + \left[1 - \delta\left(\frac{\gamma - \beta\alpha}{\gamma - \alpha}\right)\right]\gamma(1 - \gamma)^{n-1}.$$

Equation (45) has four parameters, but (46) shows that only three are identifiable. Values of α, γ, and $\delta[(\gamma - \beta\alpha)/(\gamma - \alpha)]$ are sufficient to determine the probabilities of all data.

In the form of (46), the loss of a parameter comes about because the coefficients of two terms are complementary. This may seem coincidental, but further thought shows why it must be true. Since $P(X = n)$ is a probability distribution, the sum of all its terms must equal one. This sum involves the terms

$$\sum_{n=1}^{\infty} \alpha(1 - \alpha)^{n-1}, \qquad \sum_{n=1}^{\infty} \gamma(1 - \gamma)^{n-1},$$

both of which equal one. Thus, in order that the probability distribution should sum to one, the coefficients of these two terms must also sum to one.

Of course, (45) represents only a part of what is going on in the model of (42). In addition to observing the number of trials before the first correct response, we also can observe whether an item has errors after the first correct response. This gives a probability distribution with two main components: $P(n$ errors before first correct and no errors afterward) and $P(n$ errors before first correct and one or more errors afterward). Each of these is an expression like (46) in that it has a term with $\alpha(1 - \alpha)^{n-1}$ and $\gamma(1 - \gamma)^{n-1}$; the coefficients of these terms are more complicated than in (46). In that case, the four coefficients are restricted because they must sum to one, and that restriction results in the loss of the remaining parameter.

REFERENCES

Bernbach, H. A. Derivation of learning process statistics for a general Markov model. *Psychometrika*, 1966, 31, 225–234.

Burke, C. J. and Rosenblatt, M. A Markovian function of a Markov chain. *Annals of Mathematical Statistics*, 1958, 29, 1112–1122.

Greeno, J. G. Paired associate learning with short term retention: Mathematical analysis and data regarding identification of parameters. *Journal of Mathematical Psychology*, 1967, 4, 430–472.

Greeno, J. G. Identifiability and statistical properties of two-stage learning with no successes in the initial stage. *Psychometrika*, 1968, 33, 173–215.

Greeno, J. G. and Steiner, T. E. Markovial processes with identifiable states: general considerations and application to all-or-none learning. *Psychometrika*, 1964, 29, 309–333.

Greeno, J. G. and Steiner, T. E. Comments on "Markovian processes with identifiable states: general considerations and applications to all-or-none learning." *Psychometrika,* 1968, **33**, 169–172.

Millward, R. B. Derivations of learning statistics from absorbing Markov chains. *Psychometrika,* 1969, **34**, 215–232.

Steiner, T. E. and Greeno, J. G. An analysis of some conditions for representing N state Markov processes as general all-or-none models. *Psychometrika,* 1969, **34**, 461–487.

Manuscript received 5/1/69
Revised manuscript received 1/29/71