Likelihood-based estimation and model selection for ACT-R cognitive models

Rhiannon Weaver

September 22, 2004

Abstract

ACT-R cognitive models are examples of complex hidden Markov processes with a large state space and a sparse transition matrix with absorbing states. While these models are designed to be theoretically interpretable as the thought processes generating observable data, mathematical model estimation and evaluation methods are based on simulation and prediction. There is little methodology applied to ACT-R that addresses questions of model complexity, identifiability and generalizability – all topics that can be addressed through study of the likelihood. In this paper I develop likelihood-based estimation using MCMC methods for a class of ACT-R models. I compare selection of ACT-R models for scatterplot generation using current predictive methods and using likelihood-based estimation and Bayes Factors. I propose extending this methodology to include a larger range of ACT-R models and to incorporate other model selection criteria (BIC, MDL, etc.) and decision theoretic approaches, with applications to ACT-R models including individual differences, and general hidden Markov processes.

1 Introduction

In this proposal I outline likelihood-based estimation for a constrained class of ACT-R cognitive models – models built to mimic the thought processes and serialized actions of human subjects in goal-oriented tasks. One goal of this research is to apply formal model selection criteria such as BIC, MDL and Bayes factors to ACT-R model selection problems and to evaluate the utility of such criteria in light of ACT-R modeling goals. In a broader context, however, the ACT-R architecture is an application of hidden Markov processes (Baum and Petrie, 1966) where the underlying Markov regime, posited by the model as the progression of human thought, has a large state space and a sparse transition matrix with absorbing states. In practice only a few states or a few variables from the state space can be observed. Also, the distributions of observable responses such as latency of recall and reaction times depend upon the underlying Markov process. In addition to the adaptation of model selection criteria for this framework, other questions of interest for the ACT-R applications include parameter estimation (including estimating transition probabilities for the Markov regime), efficient imputation of missing data within an MCMC framework, assessing identifiability, and measuring complexity.

Section 2 provides an overview of ACT-R history, uses and evaluation methodology, while Section 3 develops a constrained class of ACT-R models as Markov models with a large number of completely hidden states. Section 4 provides an example of model comparison as a proof of concept, and Section 5 discusses future avenues of research for the dissertation.

2 The ACT-R Cognitive Architecture

2.1 History and Use

In the middle part of the 20th century, research in cognitive psychology was conducted via "divide and conquer" techniques (Anderson and Lebiere, 1998; Introduction). In order to understand the rules and processes that comprise human thought and action, researchers focused on highly specific aspects of cognition and attempted to understand all of the details of these aspects. This methodology allowed for the study of human cognition through a series of simple experiments, and led to the discovery of many thousands of quantitative rules of human performance such as Fitt's Law [1] and the power law of practice [2].

At the 1972 Carnegie Symposium, Allen Newell began a new direction for the field of cognitive psychology. In his noted talk, "You can't play 20 questions with nature and win" (Newell, 1973a), he pushed the need for unified theories that would tie together the disparate threads of the science and, eventually, address *all* facets of cognition – problem solving, action, perception, motor, language, motivation, emotion, etc. – simultaneously (Newell, 1990; Introduction). Any such unified theory of cognition would provide a body of underlying mechanisms that, through compilation and interaction, could re-produce and expand upon previously known quantitative rules. Far more than "black boxes", these mechanisms would have theoretical interpretations. They would produce not only predictions and simulations, but also explanations, designs, and controls, providing a theoretical explanation for a phenomenon based on the functionality of the brain, in a way that a simple statistical model, like the linear model describing Fitt's Law, cannot.

In a companion paper (Newell, 1973b), Newell proposed such a system – the first production system theory of cognition – which eventually became the Soar system (Newell, 1990; Chapter 4). During the 1970's and early 1980's, Newell noted other "harbingers" of unified theories of cognition, among them the

[1] $MT \propto -\log(2D/S)$: the time MT to move to a target varies as the logarithm of the ratio of the distance D to the target and the size S of the target.

[2] $RT \propto N^{-\alpha}$: the reaction time RT for performing a task decreases as a power of the number N of times the task is practiced.

design of the *Model Human Processor* (MHP; Newell, 1990, pp 29-36), and the development of ACT* (Anderson, 1983) – which eventually evolved into ACT-R.

The MHP was designed to be a "virtual user" for programmers interested in testing human-computer interactions during software development. Its goals were far less ambitious than describing all of human cognition; instead, the model describes in detail the goal-oriented actions of a typical computer user, based on a small set of guiding principles and a few quantitative rules governing reaction time. Because the MHP is used in place of human subjects, it is important to obtain all parameter values *a priori*. Newell describes this technique as a "zero-parameter fit," but in reality, parameters are set to average values based on prior quantitative rule studies, for example Salthouse's studies on transcription typing (Salthouse, 1986). [3]

ACT* had a broader scope than the MHP, with the goal of defining a virtual human subject for a variety of experimental situations. ACT-R evolved to incorporate the rational analysis of Anderson (Anderson, 1990), and a later extension, ACT-R/PM, added auditory, visual and motor modules. With these components, ACT-R is a tool that can not only reproduce a model like the MHP, but can also generalize to the modeling of countless other goal-oriented tasks.

ACT-R models are now widely used in cognitive psychology, for more purposes than developing user simulations when data is hard to come by. ACT-R has been used in the development of cognitive tutors (Anderson, Corbett, Koedinger and Pelletier, 1995) that trace a student's knowledge of skills and strategies across a set of tasks. Moreover, ACT-R is a compelling tool for testing different symbolic representations of thought progression, each encoding the use of different strategies or skills (for example Taatgen and Anderson, 2002; Koedinger and MacLaren, 2002). Or, the underlying mechanisms themselves are estimated, as for example with terminal models (Salvucci and Anderson, 1998) and models incorporating individual differences (for instance Daily, Lovett and Reder, 2001). In these situations, experimental data is often available, and an ACT-R model is developed as a theoretical explanation of the observed data for the experiment. This is a shift in perspective; rather than a tool for producing precise simulations like the MHP, the ACT-R model is posited as directly modeling the observed human behavioral data.

2.2 Technical Overview

A basic technical overview of ACT-R follows in this section; a comprehensive overview of ACT-R can be found in Anderson and Lebiere (1998), chapters 1 through 4.

An ACT-R model consists of a symbolic level and a sub-symbolic level. At the symbolic level are a set of declarative facts, called *chunks*, each having a certain set of attributes, together with a set of procedural IF-THEN statements, called *productions*, that are used in sequence to modify chunk attributes. Figure 1 describes the flow of information within an ACT-R model. Declarative memory contains chunks that can be modified by the productions contained in procedural memory. However, a chunk can be modified only if it is in one of two retrieval buffers for the declarative memory module. The first buffer holds a chunk specifically designated as the "goal" of the task, while the second retrieval buffer can hold any chunk from declarative memory. Collectively the retrieval buffers are known as *working memory*. The IF statements in each production are conditions on the state of working memory. The system is serialized by allowing only one production to fire at any one time.

The sub-symbolic level consists of the rules of conflict resolution, recall and learning for a model. If the IF conditions for more than one production are matched, the system chooses one production to fire based on a noisy utility value for each production. Within the conflict resolution process, the model can also learn to favor certain productions over others, depending on how efficient each production is in successfully completing a goal. When no productions match the state of working memory, the model has reached an

[3] Perhaps to a Bayesian statistician these would be better named "zero-data" fits, as parameters take on well-defined prior distributions with no need for updating in light of new users, given how the model is used in practice.

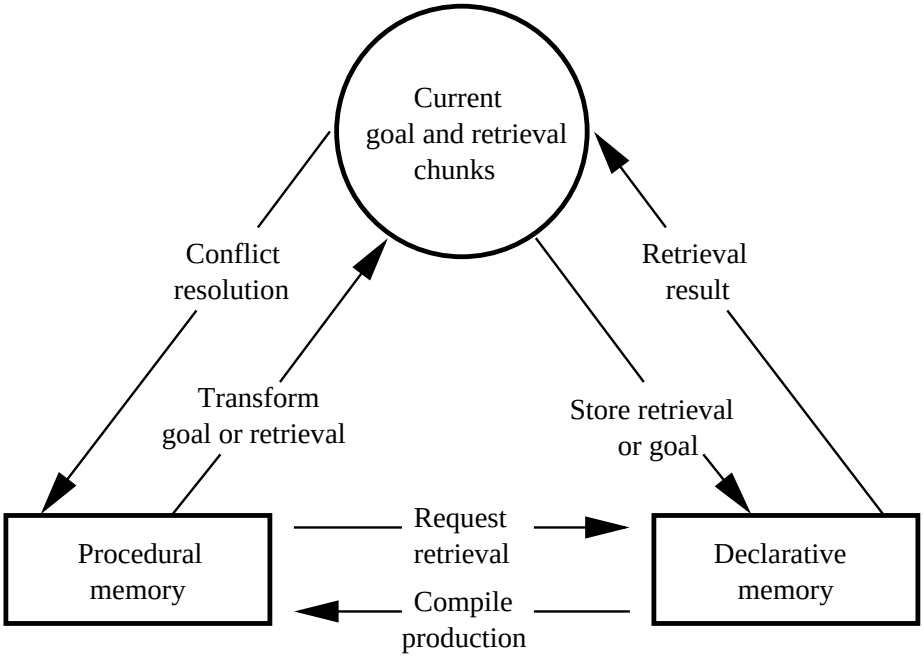

Figure 1: The flow of information in ACT-R models. Modified from Anderson and Lebiere Figure 1.2 (1998) and Byrne (2001).

end state and terminates. Productions can request retrievals from declarative memory into either the goal buffer or the retrieval buffer, and when there is more than one chunk in declarative memory that matches a production retrieval call, one chunk is chosen for retrieval based on a noisy activation value for each chunk. Information in declarative memory can also be compiled into new production rules, a form of learning.

This serialized sequence of goals, facts and IF-THEN statements is very versatile. ACT-R systems have been very successful in predicting and simulating a wide variety of human behavior. But the core of any ACT-R model can be re-envisioned as a finite state machine with an associated set of probabilities for moving from state to state defined by the rules of conflict resolution.

2.3 Current Methods

In many applications, an ACT-R model provides the specific and complicated functional form of probability for each cell in a table, where cells may refer to repeated measures across a single population (for instance, Anderson and Lebiere, 1998, chapter 4), or to a multivariate measure taken across independent populations (for instance, Baker, Corbett and Koedinger,2003). Testing theories of cognition with ACT-R involves building a model using the cognitive architecture, estimating model parameters, and comparing simulated data from the model to empirical data, often in the form of difference measures of percentages across experimental conditions. A model is selected whose percentages in each experimental condition most closely match the observed population percentages. Cell probabilities are difficult to calculate analytically, and are more often estimated via simulation.

Comparing competing explanations in ACT-R involves model estimation, evaluation and selection. The main focus in evaluation and estimation is on two distinct samples (human data and simulated). There is no direct link between the model and the human data, such as for example the likelihood of each observed data point in a generalized linear model (McCullagh and Nelder, 1989) or a generalized additive model (Buja,

Hastie and Tibshirani, 1989). ACT-R models can incorporate individual differences in parameters (eg Daily, Lovett and Reder, 2001), but in many cases the ACT-R model is used as an example of average behavior, with differences in response attributable solely to the noise in the production cycle.

2.3.1 Evaluation Criteria

Schunn and Wallach (2001) survey a number of fit statistics for cognitive and other psychological models. For K experimental conditions, all of these measures involve comparing d_k, the k-th observed percentage, and m_k, the k-th percentage as predicted by the model. The most widely used measure for evaluating the fit of the model to the trend in the data are the Pearson correlation coefficient r and the related r^2. The correlation coefficient r measures the strength of the linear relationship in the pairs (m_k, d_k), while r^2 is the proportion of variation explained by this linear relationship, assuming it to be true. These measures do not take into account the variability within groups or the uncertainty in any estimated model parameters. They can be high even when the regression line between observed and predicted values does not have slope equal to 1 and intercept 0. Also, correlation is sensitive to outliers. A condition with a very large or very small effect relative to the other conditions can have a very large effect on the correlation; the least squares line fitted through highly influential points in a regression can completely mis-represent smaller trends and still yield high r and r^2 values. Finally, for cross-classified data, different representations of the data (for example, marginal percentages and conditional probabilities vs. joint probabilities) may possibly produce different values of r or r^2.

For evaluating deviations of the model percentages m_k from the exact data percentages d_k, the root mean square deviation (RMSD), root mean square scaled deviation (RMSSD), mean absolute deviation (MAD) or mean absolute scaled deviation (MASD) are popular measures (see also Myung, 2000). Schunn and Wallach eschew traditional statistical goodness-of-fit hypothesis tests based on χ^2 statistics, noting that the null-hypothesis framework is ill-equipped for model selection, as it relies on a model being posited as true and can only provide evidence rejecting that model. Hypothesis testing for multivariate means, such as in profile analysis (Johnson and Wichern, 1998; pp.343-349), is not discussed.

ACT-R models are also evaluated on the governing principles of ACT-R theory, such as maintaining an atomic functionality of productions (Anderson and Lebiere, 1998: pp 12–13), and having components of a model (productions and chunks) also be learn-able by the model through experience (Anderson and Lebiere, 1998: p 16). These principles are more important for models that seek to explain cognitive phenomena explicitly with ACT-R architecture than for models employed as useful predictive or simulation tools.

2.3.2 Estimation, Identifiability and Complexity

ACT-R models have a number of parameters that can be estimated, though in practice only a few parameters are estimated in any one model. Usually, parameters that are estimated are global parameters such as noise parameters or threshold values (see section 3.4). Estimation methods are based upon the same objective function used to evaluate the fit of the model. Techniques such as iterated gradient descent (see Hastie, Tibshirani and Friedman, 2001; pp. 353–354) are used in some cases to find the parameters that maximize the objective function, but in many cases data-based parameter estimation is not done at all. Rather, global parameters are set to reasonable, interpretable values, and the conflict resolution parameters are learned by the model over time (Petrov, 2001).

ACT-R theory approaches identifiability by constraining existing parameters to principled, neurally plausible default values (Anderson and Lebiere, 1998, p. 17). While this is reasonable for models that produce simulations for generic experimental subjects, when the models are being used to predict the behavior of sub-populations or individuals, or when the particular interest is in the behavior of a certain parameter – for

example, exploring whether or not there are individual differences in activations or thresholds for a population – it is advantageous to make use of new observations when they are available. Furthermore, there is no quantitative theory to guarantee that parameter values selected *a priori* are the most likely values for the observed experimental data.

Though model comparison using only measures of fit is discouraged in the psychological community (Roberts and Pashler, 2000; Pitt, Myung and Zhang, 2002), there is little discussion of complexity in the ACT-R literature beyond the gold standard of the zero-parameter fit. Many statistical model selection criteria use the number of parameters as a guide, but it is still debatable what counts as a parameter in ACT-R. In practice, better zero-parameter fits are achieved by increasing the number of productions in the model (for example Byrne, 2001). But Baker Corbett and Koedinger (2003) suggest that productions should count as parameters, even when their utility values are not estimated. It is clear that adding productions to a model increases its flexibility, and with ACT-R models the functional form may be more important to model complexity than the freely varying parameters.

To the statistician, a natural place to start exploring identifiability, complexity and generalizability is with the likelihood. Likelihood-based statistical model selection criteria such as AIC (Akaike, 1973), BIC (Schwarz, 1978), and MDL (Barron, Rissanen and Yu,1998), as well as Bayesian methods like DIC (Spiegelhalter, Best, Carlin and Van der Linde, 2000) and Bayes Factors (Kass and Raftery, 1995), provide a starting point for addressing complexity in terms of the dimension of the model space and the model's functional form, identifiability in terms of the uniqueness of the likelihood surface, and generalizability in terms of uncertainty about parameter estimates and avoiding over-fitting. Using Bayesian estimation also yields posterior distributions that can be used to explore other more general utility functions.

3 Methodology for a Constrained Class of ACT-R Models

In deriving methodology for estimation for ACT-R models, I start by considering a class of ACT-R models for predicting or explaining an observable data vector X such that the conditions in Table 1 apply.

(i)	Each model has a fixed set of productions P.
(ii)	Each model has a fixed set of chunks C containing a single goal chunk.
(iii)	All chunk attributes in a model take on a finite number of discrete values
(iv)	No learning occurs (either by compiling new productions or by updating stochastic preferences).
(v)	when a model terminates, the observed data X is completely obtained from the current values of chunk attributes.

Table 1: Conditions for a constrained class of ACT-R models.

This is a highly constrained class of non-learning or "terminal" models (Salvucci and Anderson, 1998), but it provides a good starting point for discussing the statistical framework of ACT-R models.

Let $c \in C$ be a chunk with a corresponding set of attributes A_c. From (iii) above, each chunk attribute $a \in A_c$ can be envisioned as a discrete-valued categorical variable. At any time point t during the running of the model, define the current state η_t of the machine as the set of features in Table 2.

From (i) and (ii) in Table 1, the state vector η_t does not grow in dimensionality with t. From (iii) in Table 1, the space of states that the model can visit is finite, and thus η_t can be represented by a numerical value s between 1 and S, where S is the cardinality of the state space. In theory S can be very large, but in practice, the number of possible visitable states for an ACT-R model may be only a small subset of the total number of denumerable states. Models assume only a few starting states, and there are strict theoretical rules

1.	The set of all chunks and values of all chunk attributes in declarative memory.
2.	An indicator of the current chunk (ie, set of attributes) in the goal buffer.
3.	An indicator of the current chunk (ie, set of attributes) in the retrieval buffer
4.	An indicator of the most recent production that was fired.

Table 2: The elements of the current state

that limit the functions that can be encoded into productions. Using productions as the "atoms of cognition" often requires a chain of productions to fire serially (with no ability for branching) to achieve a complicated strategy or goal. Both of these properties reduce the number of visitable states.

Given a discrete time frame t that records the state η_t after each serial production cycle, the vector $\{\eta_t : t = 0, 1, ...\}$ is a random sequence of values over the space $\{1, 2, ...S\}$. Using the fixed set of productions P and conflict resolution functions, it is possible to compute the probability distribution

$$P(\eta_{t+1} = s|\eta_0, ..., \eta_t), \quad s = 1, 2, ...S. \tag{1}$$

This distribution depends on the production cycle and typically contains few non-zero probabilities. The production cycle consists of two steps:

I) determining the set of matching productions and selecting a production to fire;

II) if the matching production calls for a retrieval, determining the set of matching chunks and selecting one to retrieve into the declarative memory buffer.

Step (I) depends only on the chunk attributes in working memory in the most recent state η_t, and step (II) depends on the chunks and chunk attributes in declarative memory in the most recent state η_t. So for any state s,

$$P(\eta_{t+1} = s|\eta_0, ..., \eta_t) = P(\eta_{t+1} = s|\eta_t). \tag{2}$$

Thus $\{\eta_t : t = 0, 1, ...\}$ is a Markov chain with state space $\{1, 2, ...S\}$ and transition probabilities determined by the equations governing steps (I) and (II) in the production cycle. The Markov property is achieved by the inclusion of components 1, 2 and 3 from Table 2 in the current state. The fourth component in Table 2, an indicator for the production that fired in order to reach the current state, is needed in order to achieve a one to one map between transitions of state and production cycles.

End states in the ACT-R model correspond to absorbing states in the Markov chain. Call the set of end/absorbing states E. Suppose the model has a denumerable set of allowable starting states B. For subject i, define a *path* through the model as a vector of states $\boldsymbol{\eta}_i = (\eta_{i0}, \eta_{i1}, ...\eta_{ik})$ that trace a valid progression of production cycle transitions from $\eta_{i0} \in B$ to the value k such that $\eta_{ik} \in E$ is the first encountered end state. From condition (v), the data X that is observed for this path is a subset of the chunk attributes in η_{ik}, and the states $(\eta_{i0}, ..., \eta_{ik-1})$ are latent variables representing the progression of thought leading to X. Defining the set of states in B is a modeling decision. Starting states should be interpretable as baseline knowledge states before attempting the modeled task. A simple decision is to allow only one starting state η_0.

Let M_X be the collection of paths $\boldsymbol{\eta}_i$ such that η_{ik} produces observed data X. The probability of any response vector X can then be written as

$$P(X) = \sum_{\boldsymbol{\eta}_i \in M_X} P(\boldsymbol{\eta_i}). \tag{3}$$

The summation over all paths makes estimation of overall response probabilities unwieldy; in estimation I will focus on calculating transition probabilities from state to state.

3.1 Calculating Transition Probabilities

The state space can be explored thoroughly by starting with the common starting state η_0 and using a deterministic search algorithm, such as a depth-first search, to explore all possible states that can be reached from this initial state.

The transition probabilities are based on the components of the sub-symbolic architecture; a *utility* for each production and an *activation* for each chunk. Following Anderson and Lebiere (1998), chapters 2 and 3, the utility of a production j is given as,

$$U_j = \rho_j G - C_j, \tag{4}$$

where G is a global expected gain parameter, C_j is the cost in seconds for production j to fire, and ρ_j is the expected probability of achieving the goal given that production j fires. The parameter ρ can further be decomposed into $\rho = qr$, where q is the probability of the production firing correctly and r is the probability of a successful goal given that the production has fired. In many models, G and C_j are set to default values. In terminal models, it is of interest to estimate ρ or r (see Salvucci and Anderson, 1998).

Chunk activations function in much the same way for selecting a chunk when more than one chunk matches a current retrieval request. The activation equation for chunk ℓ takes the form

$$A_\ell(s) = b_\ell + F(\ell, s) + M(\ell, s), \tag{5}$$

where b_ℓ is a baseline activation for the chunk, $F(\ell, s)$ encodes a ***fan effect*** (Anderson and Lebiere, 1998, pp. 82-87) calculated from the state of the retrieval buffers, and $M(\ell, s)$ implements *partial matching* (Anderson and Lebiere, 1998, pp. 76-80). The functions $F(\ell, s)$ and $M(\ell, s)$ contain global parameters that are usually set to default values, but any one of the parameters could also be estimated. I restrict attention here to estimating b_ℓ.

Conflict resolution in ACT-R is performed by adding noise values $\xi_j \sim \text{logistic}(0, \sigma_u)$ to the utility values of the productions in competition and choosing the highest production to fire, with an analogous scheme for chunks, adding $\nu_\ell \sim$logistic$(0, \sigma_a)$ to activation values. The logistic distribution is used for computational efficiency during simulation in running ACT-R models, as the logistic CDF has a simple closed form. But the probability that production j fires in a set of M matching productions involves calculating $M - 1$ probability statements of the form $P(X > Y)$ when X and Y have logistic distributions. An approximate closed-form solution is

$$P_j(s) = Pr(\text{production } j \text{ fires from state } s) = \frac{e^{U_j/\sqrt{2}\sigma_u}}{\sum_{m=1}^{M} e^{U_m/\sqrt{2}\sigma_u}}. \tag{6}$$

Similarly, when a production requests a retrieval, an approximate closed-form solution for the probability that chunk ℓ will be recalled in a set of N matching chunks for a current state is

$$P_\ell(s, j) = Pr(\text{chunk } \ell \text{ recalled from state } s \text{ by production } j) = \frac{e^{A_\ell(s)/\sqrt{2}\sigma_a}}{\sum_{n=1}^{N} e^{A_n(s)/\sqrt{2}\sigma_a}}. \tag{7}$$

See Anderson and Lebiere (1998; Appendix A, pp. 89–92) for details on these approximations.

Each transition of state is achieved by selecting a production using the rule defined by equation 6, and if a retrieval is called, selecting a chunk using the rule defined by equation 7. Including an indicator in the current state for the most recent production that fired creates a one-to-one map between state transitions and production cycles. Specifically, for any two non-identical states $s_1, s_2 \in 1, ..., S$, there exists at most one

distinct production cycle that transitions s_1 directly to s_2. In this case, transition probabilities all have the general form

$$P_j(s) \times [P_\ell(s,j)]^z, \tag{8}$$

for some production j and retrieval chunk ℓ, where $P_j(s)$ and $P_\ell(s,j)$ are defined in Equations 6 and 7 and z is equal to 0 if production j does not request a retrieval, and 1 otherwise.

3.2 Functional form of the Likelihood

The joint likelihood of an observed response X together with the subject's path through S is the product of transition probabilities leading to the end state. Formally, suppose there exist I independent subjects and an ACT-R model whose state space consists of S states, with L chunks and corresponding vector $\mathbf{b}$ of baseline activation parameters, and J productions with corresponding vector $\boldsymbol{\rho}$ of utility parameters. Define

$\boldsymbol{\eta_i}$: a vector of length K_i representing subject i's path, where each η_{ik} is a numerical value between 1 and S.

ϵ: a $J \times S$ indicator matrix where ϵ_{js} is 1 if production j matches state s and 0 otherwise.

δ: a $L \times J \times S$ indicator matrix where δ_{ljs} is 1 if chunk l can be recalled by production j in state s, and 0 otherwise.

Y_i: a $J \times K_i$ indicator matrix where Y_{ijk} is 1 if production j fired on the k-th step of subject i's path, and 0 otherwise.

Z_i: a $L \times K_i$ indicator matrix where Z_{ilk} is 1 if chunk l was recalled on the k-th step of subject i's path, and 0 otherwise.

Note that δ and ϵ are determined solely by the structure of the ACT-R model and are not estimated parameters. Also by condition (iv) in the class of ACT-R models, the observed data X_i for subject i is a subset of the information in state η_{iK_i}. The likelihood can be written as

$$\begin{aligned} L(\mathbf{X},\mathbf{Y},\mathbf{Z},\boldsymbol{\eta}|\boldsymbol{\rho},\mathbf{b},\sigma_a,\sigma_u) &= \prod_{i=1}^{I}\prod_{k=1}^{K_i}\prod_{j=1}^{J}\left[\frac{\exp(U_j/\sqrt{2}\sigma_u)}{\sum_{m=1}^{J}\epsilon_{m\eta_{ik}}\exp(U_m/\sqrt{2}\sigma_u)}\right]^{Y_{ijk}} \\ &\times \prod_{l=1}^{L}\left[\frac{\exp(A_l(\eta_{ik})/\sqrt{2}\sigma_a)}{\sum_{n=1}^{L}\delta_{nj\eta_{ik}}\exp(A_n(\eta_{ik})/\sqrt{2}\sigma_a)}\right]^{Z_{ilk}}, \end{aligned} \tag{9}$$

The ϵ and δ coefficients select the relevant productions or chunks to include in the denominator for each production cycle, producing the transition probabilities $P_j(\eta_{ik}) \times [P_l(j,\eta_{ik})]^z$ as described in section 3.1. Due to the one-to-one mapping between production cycles and state to state transitions, the indicators $\mathbf{Y}$ and $\mathbf{Z}$ can be completely reconstructed from the information in the path vectors $\boldsymbol{\eta_i}$ for each subject. From now on I will suppress $\mathbf{Y}$ and $\mathbf{Z}$ in joint and conditional expressions.

It is advantageous to think of the entire path vector η_i as a latent indicator in a mixture model (Gelman, Carlin, Stern and Rubin, 1995; chapter 16), where the mixing distribution describes the prior distribution over paths (uniform as described by Equation 3), and the functional form of the likelihood changes from path to path according to the indicator matrices ϵ and δ, while relying on subsets of the same collection of utility and activation parameters. Posterior inference on utility, activation and noise parameters will be obtained by data augmentation (Tanner and Wong, 1987) within a Gibbs sampler (Gelman, Carlin, Stern and Rubin, 1998, chapter 11; Tanner, 1996, chapter 6).

3.3 An MCMC Algorithm for Estimation

For ACT-R models, a Gibbs sampler using data augmentation alternates between (i) sampling a path vector for each subject conditional on the observed data for that subject and the current values of the model parameters and (ii) updating model parameters conditional on the path for each subject.

3.3.1 Sampling Paths

The likelihood of a path $\boldsymbol{\eta_i}$ for subject i is given by

$$\begin{aligned} L(X_i, \boldsymbol{\eta}|\boldsymbol{\rho}, \mathbf{b}, \sigma_a, \sigma_u) &= \prod_{k=1}^{K_i} \prod_{j=1}^{J} \left[\frac{\exp(U_j/\sqrt{2}\sigma_u)}{\sum_{m=1}^{J} \epsilon_{m\eta_{ik}} \exp(U_m/\sqrt{2}\sigma_u)} \right]^{Y_{ijk}} \\ &\times \prod_{l=1}^{L} \left[\frac{\exp(A_l(\eta_{ik})/\sqrt{2}\sigma_a)}{\sum_{n=1}^{L} \delta_{nj\eta_{ik}} \exp(A_n(\eta_{ik})/\sqrt{2}\sigma_a)} \right]^{Z_{ilk}}. \end{aligned} \tag{10}$$

Assuming a uniform prior on paths, the conditional distribution of paths given the set of model parameters is proportional to this likelihood when it is non-zero. The observed data X restricts the set of paths with non-zero probability to those in the set M_X, but selecting a candidate path directly from M_X by enumerating all paths can be combinatorially difficult. One can employ ACT-R simulation in a form of rejection sampling (Gelman, Carlin, Stern and Rubin; 1995). In this situation the target distribution $p(\eta|X)$ is the distribution normalized over paths in M_X with zero probability elsewhere, and the proposals are drawn from the joint distribution $q(\eta, X)$ of all possible paths and responses. This method can be easily implemented with existing ACT-R software, however depending on the size of M_X with respect to the set of all paths, and the relative likelihood of the observed data, it could be very inefficient.

To avoid enumerating the set M_X explicitly, it is sufficient to find the subset E_X of states in the set of end states E that match the response vector. Once a viable end state is chosen from E_X, an entire path can be sampled using the transpose of the transition matrix to obtain valid next-state candidates, and completed when the common starting state η_{i0} is reached.

Another alternative is to sample an entire path backward from an end state using a simple discrete proposal distribution, and then to perform a Metropolis step. Let θ_t be the set of model parameters at iteration t, $\boldsymbol{\eta_i}^t$ the path at iteration t, and suppose J_t is the proposal (or "jumping") distribution used to sample a path. The Metropolis step accepts a candidate path $\boldsymbol{\eta_i}^*$ with probability,

$$r = \min\left(\frac{L(\boldsymbol{\eta_i}^*|X, \theta_t) J_t(\boldsymbol{\eta_i}^t|\boldsymbol{\eta_i}^*, \theta_t)}{L(\boldsymbol{\eta_i}^t|X, \theta_t) J_t(\boldsymbol{\eta_i}^*|\boldsymbol{\eta_i}^t, \theta_t)}, 1 \right).$$

The simplest form for J_t is a conditional uniform distribution; at each branching point a candidate state is chosen uniformly out of all possible candidates. This proposal distribution is independent of the current path and the model parameters, and so the Metropolis calculation reduces to

$$r = \frac{L(\boldsymbol{\eta_i}^*|X, \theta_t) J_t(\boldsymbol{\eta_i}^t)}{L(\boldsymbol{\eta_i}^t|X, \theta_t) J_t(\boldsymbol{\eta_i}^*)}. \tag{11}$$

This is not the most efficient form of proposal, but it is easy to calculate within the Gibbs sampler. However for sets M_X with a very large number of possible paths or with sets of very unlikely end states, it could be inefficient.

3.3.2 Updating Model parameters

Let $\pi(\boldsymbol{\rho}), \pi(\mathbf{b}), \pi(\sigma_u)$ and $\pi(\sigma_a)$ be independent prior distributions on utility, activation and noise parameters. The posterior $p(\boldsymbol{\rho}, \mathbf{b}, \sigma_a, \sigma_u | \mathbf{X}, \boldsymbol{\eta})$ is proportional to

$$p(\boldsymbol{\rho}, \mathbf{b}, \sigma_a, \sigma_u | \mathbf{X}, \boldsymbol{\eta}) \quad \alpha \quad L(\mathbf{X}, \boldsymbol{\eta} | \boldsymbol{\rho}, \mathbf{b}, \sigma_a, \sigma_u) \pi(\boldsymbol{\rho}) \pi(\mathbf{b}) \pi(\sigma_u) \pi(\sigma_a) \tag{12}$$

For a utility parameter ρ_j, the complete conditional distribution is proportional to

$$p(\rho_j | \mathbf{X}, \boldsymbol{\eta}, \boldsymbol{\rho}_{-j}, \mathbf{b}, \sigma_a, \sigma_u) \quad \propto \quad \prod_{i=1}^{I} \prod_{k=1}^{K_i} \left[\frac{\exp(U_j / \sqrt{2}\sigma_u)}{\sum_{m=1}^{J} \epsilon_{m\eta_{ik}} \exp(U_m / \sqrt{2}\sigma_u)} \right]^{Y_{ijk}} \times \pi(\rho_j). \tag{13}$$

For an activation parameter b_l, the complete conditional distribution is proportional to

$$p(b_l | \mathbf{X}, \boldsymbol{\eta}, \boldsymbol{\rho}, \mathbf{b}_{-l}, \sigma_a, \sigma_u) \quad \propto \quad \prod_{i=1}^{I} \prod_{k=1}^{K_i} \prod_{j=1}^{J} \left[\frac{\exp(A_l(\eta_{ik}) / \sqrt{2}\sigma_a)}{\sum_{n=1}^{L} \delta_{nj\eta_{ik}} \exp(A_n(\eta_{ik}) / \sqrt{2}\sigma_a)} \right]^{Z_{ilk}} \times \pi(b_l) \tag{14}$$

For the utility noise parameter σ_u, the complete conditional is proportional to

$$p(\sigma_u | \mathbf{X}, \boldsymbol{\eta}, \boldsymbol{\rho}, \mathbf{b}, \sigma_a,) \quad \propto \quad \prod_{i=1}^{I} \prod_{k=1}^{K_i} \prod_{j=1}^{J} \left[\frac{\exp(U_j / \sqrt{2}\sigma_u)}{\sum_{m=1}^{J} \epsilon_{m\eta_{ik}} \exp(U_m / \sqrt{2}\sigma_u)} \right]^{Y_{ijk}} \times \pi(\sigma_u), \tag{15}$$

and for the activation noise parameter σ_a, the complete conditional is proportional to

$$p(\sigma_a | \mathbf{X}, \boldsymbol{\eta}, \boldsymbol{\rho}, \mathbf{b}, \sigma_u) \quad \alpha \quad \prod_{i=1}^{I} \prod_{k=1}^{K_i} \prod_{j=1}^{J} \prod_{l=1}^{L} \left[\frac{\exp(A_l(\eta_{ik}) / \sqrt{2}\sigma_a)}{\sum_{n=1}^{L} \delta_{nj\eta_{ik}} \exp(A_n(\eta_{ik}) / \sqrt{2}\sigma_a)} \right]^{Z_{ilk}} \times \pi(\sigma_a). \tag{16}$$

Distributions of noise parameters are similar to those of their associated utility (or activation) parameters, except since noise parameters are the same across all productions (or chunks) the product in the likelihood is taken across all productions (or across all chunks).

The conditional distribution of a model parameter is always proportional to the product of the prior and the terms in the likelihood containing that parameter. Although the conditional distributions of model parameters do not have a closed form, it is relatively straightforward to use a Metropolis step within the Gibbs sampler to obtain draws from these distributions.

3.4 Utility and Retrieval Thresholds

In most ACT-R models, productions compete not only among each other but also against a *utility threshold* value, τ_u. A production j selected from a set of competing productions fires only if

$$U_j + \xi_j > \tau_u,$$

otherwise the model terminates. Similarly for activation, chunks compete against an *activation threshold*, τ_a. A chunk ℓ selected from a set of competing chunks fires only if

$$A_\ell(\eta) + \nu_\ell > \tau_a,$$

otherwise the model terminates. In these extensions, even productions and chunks that have no competitors (and otherwise would fire with certainty), are compared against threshold values. The utility threshold behaves exactly like an extra, virtual production that matches every state. I will designate this production as production 0, such that

$$U_0 = \tau_u.$$

Similarly, the activation threshold behaves like an extra, virtual chunk, that matches every state. I will designate this chunk as chunk 0, with activation

$$A_0(\eta) = \tau_a,$$

for all states η. Complete conditional distributions for τ_u and τ_a have the form of Equations 13 and 14.

To incorporate thresholds in path sampling, every state in the space is included in the set E of possible end states. If a state η_{iK_i} for subject i has a non-null set of matching productions, an extra term is added to the path likelihood to account for a threshold failure. With the increased number of possible end states for each subject, it is advantageous to incorporate the current state of the model parameters into the proposal distribution for paths. Let η be a candidate end state chosen for subject i. Suppose at iteration t in the Gibbs sampler, N_η out of the I total subject paths end in this state. An estimate of $P(\eta)$ is

$$\hat{P}(\eta|\theta) = \frac{N_\eta}{I}.$$

In practice there are many more end states than subjects in a study. It is advantageous to instead calculate

$$\tilde{P}(\eta|\theta) = \frac{N_\eta + \iota}{I + S \cdot \iota},$$

where ι acts like a variance tuning parameter for a proposal distribution. Let $T(\eta)$ be the event that the model terminates in state η. Define the proposal probability of state η as

$$J_t(\eta|\theta) \quad = \quad \frac{\tilde{P}(\eta|\theta)P(T(\eta)|\eta,\theta)}{\sum_{\nu \in E_X}[\tilde{P}(\nu|\theta)P(T(\nu)|\eta,\theta)]}$$

where θ is the vector of model parameters. The term $P(T(\eta)|\eta,\theta)$ sums the probabilities of utility threshold failure and activation threshold failure conditional on state η, and has the form

$$\begin{aligned} P(T(\eta)|\eta,\theta) &= 1_{\{\sum_{j=1}^J \epsilon_{j\eta}=0\}} + 1_{\{\sum_{j=1}^J \epsilon_{j\eta}>0\}} \frac{\exp(U_0/\sqrt{2}\sigma_u)}{\sum_{j=0}^J \epsilon_{j\eta}\exp(U_j/\sqrt{2}\sigma_u)} \\ &+ \sum_{j=1}^J \left(\epsilon_{j\eta} 1_{\{\sum_{l=1}^L \delta_{lj\eta}>0\}} \left[\frac{\exp(U_j/\sqrt{2}\sigma_u)}{\sum_{j=0}^J \epsilon_{j\eta}\exp(U_j/\sqrt{2}\sigma_u)} \right] \left[\frac{\exp[A_0(\eta)/\sqrt{2}\sigma_a]}{\sum_{l=0}^L \delta_{lj\eta}\exp[A_l(\eta)/\sqrt{2}\sigma_a]} \right] \right). \end{aligned} \tag{17}$$

The specific failure event F can be chosen conditional on the end state η with probability

$$\frac{P(F|\eta,\theta)}{P(T(\eta)|\eta,\theta)}.$$

Let r be the ratio using uniform jumping paths as described in Equation 11. At time t with proposal path $\boldsymbol{\eta_i}^*$, current path $\boldsymbol{\eta_i}^t$ and model parameters θ_t, the acceptance ratio r_f incorporating threshold failures is

$$\begin{aligned} r_f &= \frac{L(\boldsymbol{\eta_i}^*|X_i,\theta_t)P(F^*|\eta^*_{iK_i},\theta_t)}{|E_{X_i}|J_t(\boldsymbol{\eta_i}^*)\frac{\tilde{P}(\eta^*_{iK_i})P(T(\eta^*_{iK_i})|\eta^*_{iK_i},\theta_t)}{\sum_{\nu\in E_{X_i}}[\tilde{P}(\nu)P(T(\nu)|\nu,\theta_t)]}\times\frac{P(F^*|\eta^*_{iK_i},\theta_t)}{P(T(\eta^*_{iK_i})|\eta^*_{iK_i},\theta_t)}} \\ &\times\frac{|E_{X_i}|J_t(\boldsymbol{\eta_i}^t)\frac{\tilde{P}(\eta^t_{iK_i})P(T(\eta_{iK_i})^t|\eta^t_{iK_i},\theta_t)}{\sum_{\nu\in E_{X_i}}[\tilde{P}(\nu)P(T(\nu)|\nu,\theta_t)]}\times\frac{P(F^t|\eta^t_{iK_i},\theta_t)}{P(T(\eta_{iK_i})^t|\eta^t_{iK_i},\theta_t)}}{L(\boldsymbol{\eta_i}^t|X_i,\theta_t)P(F^t|\eta^t_{iK_i},\theta_t)}. \end{aligned}$$

This expression reduces to

$$r_f = r\times\frac{\tilde{P}(\eta^t_{iK_i})}{\tilde{P}(\eta^*_{iK_i})}$$

3.5 Experimental Conditions

Chunk attributes can be classified as either *system variables* – attributes for which there exists at least one production in the ACT-R model that is able to modify the attribute's value, or *experimental conditions* – attributes pre-set before running the model, whose values cannot be modified by ACT-R productions, but may be called as conditions in IF statements. Both the outcome X and the attributes making up the experimental conditions can be considered observable. While X can be calculated as a subset of the attributes in the terminating state of the ACT-R model, the experimental conditions are set at the starting state and fixed throughout any subject's path. This expands the set of starting states for the model. The smallest set B of starting states can be enumerated $\{\eta_0^1,...\eta_0^C\}$, where C is the number of experimental conditions.

Because the experimental conditions are not changed by productions, the transition matrix including experimental conditions is partitionable into a group of C sub-matrices. Experimental conditions can be implemented with C separate ACT-R models, each tracing the production cycles and paths expanding from a starting state η_0^c. In this case the experimental conditions do not need to be included in the current state; instead productions that can never be called starting from state η_0^c are simply not included in the determination of the c-th sub-model's state space and transition matrix.

3.6 Model Comparison Using Bayes Factors

The Bayes Factor (Kass and Raftery, 1995) for model M_1 in relation to model M_0 is the ratio of posterior to prior odds

$$B_{10} = \frac{P(X|M_1)}{P(X|M_0)},$$

which is the ratio of marginal likelihoods,

$$P(X|M_k) = \int L(X|\theta_k,M_k)P(\theta_k|M_k)d\theta_k,$$

where θ_k is the set of parameters for model k, $k=0,1$.

For an ACT-R model with parameter vector θ, the data augmentation method produces samples from the joint posterior distribution of model parameters and paths, $Pr(\theta,\eta|X)$. Kass and Raftery suggest a number

of importance sampling estimators of the marginal likelihood; the easiest one to implement using a sample from the posterior is

$$\hat{P}(X|M_k) = \left\{ \frac{1}{T} \sum_{i=1}^{T} L(X|\theta_k^i, \eta_k^i)^{-1} \right\}^{-1},$$

the harmonic mean of the likelihood values calculated at each iteration. Though it is unstable due to occasional values of (θ_k^i, η_k^i) with very small likelihood (and a large effect on the final result), it often gives enough information for interpretation on a log scale. To obtain the posterior probability of model k in relation to $K > 2$ competing models, Bayes factors can be calculated against a reference model M_0 such that

$$P(M_k|X) = \frac{\alpha_k B_{k0}}{\sum_{r=0}^{K-1} \alpha_r B_{r0}},$$

where α_k is the prior odds of model k versus model 0 (in many cases where there is no prior information about models, α_k is set to 1).

4 Model Comparison Example

4.1 Description and Data

Scatterplots display the relationships between two quantitative variables by plotting (X, Y) pairs on a Cartesian graph. Baker, Corbett and Koedinger (2003) note that in creating scatterplots, students make two kinds of errors involving the conceptual understanding of categorical versus quantitative variables. They denote these conceptual errors as *variable choice errors* and *nominalization errors*.

To explore this phenomenon, 132 middle school students were given a scatterplot task with two quantitative variables, the age of a range of singers (AGE) and the number of pieces of fan mail received by each singer (PIECES), along with the categorical variable of whether each singer was also a musician (MUSICIAN). Percentages of errors are shown in Table 3. Students who committed a variable choice error used the categorical variable MUSICIAN on one of the axes in the plot. Students who committed a nominalization error used both quantitative variables AGE and PIECES, but treated one or both as nominal variables instead of quantitative variables.

Baker, Corbett and Koedinger wished to determine whether these similar conceptual errors could be explained better by the execution of a single strategy or the execution of multiple strategies producing different results. Five experimental conditions were employed, relating to the kind of information each subject was given at the start of the assignment. These are described in Table 3. Five different models were constructed. Model KNOW-IT-ALL allowed for the understanding of both the correct representation of quantitative variables and familiarity with scatterplots. Model KNOW-SCATTERPLOTS allowed for a familiarity with scatterplots but no understanding of quantitative variables outside that context. Model KNOW-QUANTITATIVES allowed for an understanding of quantitative variables but a lack of familiarity with scatterplots. Model CAN'T-USE-QUESTION restricted the information that a student processed when presented with the task. Finally, model DON'T-KNOW-BARGRAPHS restricted nominalization errors to random guessing, rather than an inappropriate knowledge transfer from existing skills in creating bar graphs.

All models were terminal models, each consisting of subsets of a common set of 32 productions and 7 chunks. An iterated gradient descent estimation scheme was used, based on the models' predictive accuracy toward matching the percentages in Table 3, as measured by the r^2 metric described in Section 2.3.1. Between 4 and 6 utilities were estimated for each model, as well as a common baseline activation value for all chunks, the activation and utility thresholds, and the utility noise. Model selection was performed by

		No prompts	No variables labeled	X variable labeled	Y variable labeled	Both variables labeled
(i)	Sample size	13	30	29	29	31
(ii)	Variable choice error	15.0	26.9	7.7	26.9	6.5
(iii)	Correct axis variables (CAV)	0	73.1	79.3	73.1	77.4
(iv)	Given CAV, X axis nominalized only	n/a	15.7	17.4	15.7	12.5
(v)	Given CAV, Y axis nominalized only	n/a	0	0	0	0
(vi)	Given CAV, both axes nominalized	n/a	5.3	8.7	0	8.3
(vii)	Given CAV, correct representation on both axes	n/a	73.7	73.9	84.3	79.2

Table 3: Percentages of different behaviors in scatterplot construction (Baker, Corbett and Koedinger, 2003). Columns show experimental conditions, while rows list the number of students in each condition, and the observed percentages of each response. Rows (iv) through (vii) are percentages conditional on the subpopulations that labeled the axes correctly (row (iii)).

Baker et.al. using the least squares criterion,

$$LS = K \sum_{k=1}^{K} (m_k - d_k)^2 + p \log K,$$

where m_k and d_k are the model prediction and observed k-th data value from Table 3, respectively, and p is a complexity measure based on the number of parameters allowed to vary, and the number of productions and chunks in the model.

4.2 Preliminary Results

Table 4 shows an exploratory summary of the complexity of the five models estimated. Of the five models, KNOW-QUANTITATIVES has a dramatically smaller state space, even though more parameters are estimated for this model than model CAN'T-USE-QUESTION. Also, although the state space is the largest for model KNOW-IT-ALL, model DON'T-KNOW-BARGRAPHS is the only model that provides valid paths for all observed data. Table 5 shows the student IDs whose responses for each model could not be reproduced by the model. The set of matching end states for these students was empty.

Model	# Productions	# System Variables	S	# Parameters Estimated
KNOW-IT-ALL	31	13	2772	8
KNOW-SCATTERPLOTS	29	13	2526	7
KNOW-QUANTITATIVES	27	13	1713	7
CAN'T-USE-QUESTION	27	13	2670	6
DON'T-KNOW-BARGRAPHS	26	13	2635	8

Table 4: An exploratory look at the complexity of five models created in ACT-R, used to model strategy choice in scatterplot construction.

Model	Student IDs
KNOW-IT-ALL	17,26,80,107
KNOW-SCATTERPLOTS	17,26,80,107
KNOW-QUANTITATIVES	17,26,80,86,107
CAN'T-USE-QUESTION	17,26,80,107
DON'T-KNOW-BARGRAPHS	none

Table 5: Student IDs outside the model range for the five models.

For likelihood-based estimation, the chunk threshold and noise values were set to the best fit values from the iterated gradient descent performed by Baker,Corbett and Koedinger for each model. Only utility parameters, baseline chunk activation and retrieval thresholds were estimated for each model. Estimating both a retrieval threshold and a baseline activation leads to a multiplicative non-identifiability in the probability of retrieval. Noise values were kept constant in order to compare likelihood estimates for the utility and activation parameters with the predictive method.

For the utility and baseline activation parameters, Normal(0, $\sigma^2 = 100$) priors were used. For the utility threshold a Normal(0,$\sigma^2 = 10$) prior was used. G and C_j for production utilities were set to ACT-R default

values, and the noise value ι for sampling paths was set to 10. Markov chains of 950 iterations were run for each model, with 300 iterations removed as burn-in. The small number of iterations is due to the prototype software implementation in S-plus rather than a faster language such as C or C++. Inspection by eye showed quick convergence for parameters that appeared in many paths, but for parameters that were rarely used or that estimated tail probabilities, the chains were erratic.

Model	LS	$\frac{1}{T}\sum_{t=1}^{T} L(X\|\theta_k^t, \eta_k^t)$	$\log(B_{0k})$
KNOW-IT-ALL	194.1	-1365.4	341
KNOW-SCATTERPLOTS	208.2	-1033.5	0
KNOW-QUANTITATIVES	181.8	-1175.5	171
CAN'T-USE-QUESTION	245.8	-1107.0	84
DON'T-KNOW-BARGRAPHS	215.9	-1338.4	309

Table 6: Average likelihood and Bayes factors for five models of scatterplot generation. Bayes factors are taken as $\log B_{0k}$ where model 0 is model KNOW-SCATTERPLOTS.

Results of comparisons with Bayes factors using the harmonic mean estimator of the marginal likelihood are shown in Table 6. Although the predictive analysis selects model KNOW-QUANTITATIVES, the likelihood-based analysis is strongly in favor of model KNOW-SCATTERPLOTS. There is little difference in the predictive power of models KNOW-IT-ALL, KNOW-QUANTITATIVES and KNOW-SCATTERPLOTS, but model KNOW-SCATTERPLOTS achieves a good fit to the data with shorter paths, and thus less complexity.

Because longer paths multiply more terms into the likelihood, likelihood-based estimation favors shorter paths and paths with fewer branches (and thus simpler strategies), when all other factors are equal. The quantitative strategy in models KNOW-IT-ALL and KNOW-QUANTITATIVES used three productions in a row as opposed to the scatterplot strategy that used only one production. Both quantitative and scatterplot strategies yielded the same end result, with the only distinction in the likelihood being the larger probability of a threshold failure along the longer (and thus more complicated) quantitative strategy.

Analysis with Bayes factors suggests that model KNOW-SCATTERPLOTS adequately accounts for threshold failures, without the extra explanation of exploring an alternate and more difficult strategy. But in this case, the data do not provide much information for distinguishing between strategies. More information about the use of the more difficult quantitative strategy could be gained through recording reaction times, in which case a longer reaction time would give evidence of a more difficult strategy, or through a talk-aloud approach that could explicitly record a student's attempt at a quantitative approach.

5 Future Work

The scatterplot example in Section 4 presents some preliminary results and a proof of concept for the methodology developed in Section 3, but it also raises many questions for researchers interested in using ACT-R models to account for cognitive processes. Likelihood-based estimation and model selection methodology provides a framework for discussing uncertainty, identifiability and complexity as well as goodness of fit, and it can produce dramatically different results and conclusions for models that provide similar simulation-based predictions. But the usefulness of this kind of evaluation depends also on the goals of the modelers. For instance, it is perhaps more important to determine the falsifiability or interpretation of a model in theory testing than in predictive applications. And while penalty terms account for generalizability in adaptation to new data for the same task, ACT-R models are often built to generalize toward learning new tasks – in which case more same-task complexity is required.

In addition to providing likelihood-based selection criteria for predictive methods, a Bayesian estimation scheme opens the door for decision theoretic model selection based on general utility functions, and exploring estimation, identifiability and complexity for ACT-R applications provides a concrete example of implementation and feasibility for complex hidden Markov processes. Toward these goals I propose the following agenda for the dissertation:

Expand activation equations: For a chunk ℓ being called by production j in state η_{ik}, the activation equation can be expanded as

$$A_\ell(\eta_{ik}, j) = \beta - d \log t_\ell + \sum_{m=1}^{n} \frac{W}{n} S_{\ell m} - D_{\ell j},$$

where d is a decay rate, t_ℓ is the amount of time since chunk ℓ was last recalled, and m sums over the attributes in the goal chunk. Further equations govern latency of response times. The parameter W measures the amount of attention paid to each goal attribute, and $S_{\ell m}$ is a fan effect that measures the association between chunk ℓ and the state of the current goal. The partial matching penalty $D_{\ell,j}$ is based on the retrieval constraints set by production j. Any of these parameters can be estimated from the data using the general MCMC methodology outlined for activation parameters in Section 3. Implementation of these extensions in the estimation software is a straightforward programming task. Further development involves incorporating latency and reaction times.

Incorporate learning and extend observable data: Extending the methodology to incorporate learning via parameter updating is similar to the extension of static latent class models (Langeheine and Rost, 1988) to dynamic hidden Markov models (Baum and Petrie, 1966). A more complicated extension is involved for models that create new productions and chunks with repeated trials; in this case the state space and transition matrix need to be updated with each trial to include new productions and chunks.

When data X_i is retrieved entirely from the state η_{iK_i}, the set M_{X_i} of paths with non-zero probability can be explored by finding all possible end states that match X_i. On the other extreme, if entire path is made visible, as for example with intelligent tutors, the set M_{X_i} consists of only a single path. Along this continuum, sampling entire paths with non-zero probability of producing X_i is computationally more difficult than finding the set of end states, as observable data restricts matching states at more points along the path. This limited observability can be seen as a Markov process (or depending on the nature of observable data, a hidden Markov process) with missing data or gaps between observable values. One approach to this problem is to start by adapting traditional estimation techniques for hidden Markov regimes (eg. Lindgren, 1978; Holst and Lindgren, 1991) to ACT-R models, and then to expand the framework to incorporate missing data.

Calculate traditional model selection criteria for ACT-R models: In addition to Bayes Factors, Table 7 lists three statistical model selection criteria that can be applied to ACT-R models. The BIC, DIC and MDL all involve a marginal maximum likelihood calculation and a calculation of the effective number of parameters p for the model. Raftery (1995) describes techniques for maximum likelihood estimation using Monte Carlo simulations, as well as a data augmentation estimator that produces consistent estimation of Bayes factors when many nuisance parameters are present in all models under consideration.

Moody (1992) discusses the effective number of parameters for models based on least squares estimation, while Ye (1998) offers a method for calculating the effective degrees of freedom of algorithmic models, based on simulation. For hierarchical normal linear models, Ye's algorithmic approach coincides with the constraint-case method of counting parameters outlined by Hodges and Sargent (2001), also closely related to smoothers (Buja, Hastie and Tibshirani, 1989). Lee and Nelder (1996) also derive a similar method for hierarchical generalized linear models, based on a first-order approximation. The MDL also requires the integral over the parameter space of the Fisher information for a sample of size 1 (see Pitt, Myung and Zhang,

Name	Formula
BIC	$-2\log f(X\|\hat{\theta}) + p\log n$
MDL	$-\log f(X\|\hat{\theta}) + \frac{p}{2}\log\frac{n}{2\pi} + \log\int\|I(\theta)\|d\theta$
DIC	$\frac{-2}{T}\sum_{t=1}^{T}\log f(X\|\theta^t) + 2\log f(X\|\tilde{\theta})$

Table 7: Common statistical model selection criteria: Bayesian Information Criterion (BIC), Minimum Description Length (MDL), and Deviance Information Criterion (DIC). Here θ is the vector of model parameters, with $\hat{\theta}$ the maximum likelihood estimator and $\tilde{\theta}$ any Bayesian estimator such as the posterior mean or mode. The function $f(X|\theta)$ is the likelihood function, $I(\theta)$ is the Fisher Information matrix of a sample of size 1, n is the sample size, and p is the effective number of parameters in the model.

2002). But estimation may be feasible, for example using the missing information principle (Tanner, 1996, pp. 74–75.). In addition, Ito (1992) explores identifiability, complexity and information for parameters in hidden Markov regimes.

Apply methodology to existing ACT-R models: Some applications that focus on production branching and strategy choice include task switching (Anderson, Taatgen, and Byrne, 2004) and categorization (Anderson and Betz, 2001). The task switching application additionally offers the opportunity to study both a terminal model and a learning model.

While the current estimation methodology is developed for production branching as subjects choose different strategies, a large number of ACT-R models are also based upon reaction times and latency data. These models often have a common, simpler production structure for all subjects, and focus on the structure of chunk activation functions. An example is list memory (Anderson, Bothell, Lebiere, and Matessa, 1998); although based on an early version of ACT-R, the list memory model is a candidate for translation into the estimation architecture. Additionally, Daily, Lovett and Reder (2001) outline a theory of working memory for ACT-R that attributes differences in recall among subjects to individual differences in the magnitude of the weight W associated with fan effects. The working memory study offers the opportunity to compare a simple kind of nested structure in ACT-R – models with the same production structure that differ only in the expansion of a parameter to include individual differences.

Program a computationally efficient estimation algorithm: This incorporates both a translation of the existing code to C++ for faster and more portable machinery, as well as adapting the MCMC methodology for path sampling using the complete conditional distributions for individual states in the path.

6 References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. pp 267–281 in Petrov, B. N, and Csaki, F. (Eds) *Second international symposium on information theory*. Budapest: Academiai Kiado.

Anderson, J. (1983). *The Architecture of Cognition.* Cambridge, MA. Harvard University Press.

Anderson, J. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ. Lawrence Erlbaum Associates.

Anderson, J., and Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin and Review*, vol. 8, pp. 629–647.

Anderson, J., Bothell, D., Lebiere, C. and Matessa, M. (1998) List memory. Chapter 7 in Anderson, J. and Lebiere, C. (1998). *The Atomic Components of Thought.* Lawrence Erlbaum Associates.

Anderson, J. and Lebiere, C. (1998). *The Atomic Components of Thought.* Hillsdale, NJ. Lawrence Erlbaum Associates.

Anderson, J. and Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, vol. 104, pp. 728–748.

Anderson, J., Taatgen, N., and Byrne, M. (2004). Learning to achieve perfect time sharing: architectural implications of Hazeltine, Teague, and Ivry (2002). Submitted.

Baker, R., Corbett, A., Koedinger, K. (2003). Statistical techniques for comparing ACT-R models of cognitive performance. *Proceedings of the 10th Annual ACT-R Workshop*, pp. 129–134.

Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, vol. 4, no. 6, pp. 2743–2760.

Baum, L. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state machines. *Annals of Mathematical Statistics*, vol 37, pp. 1554–1563.

Breiman, L (2001). Statistical modeling: the two cultures. *Statistical Science*, vol. 16, no. 3, pp. 199–231.

Byrne, M. (2001). ACT-R/PM and menu selection; Applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies*, vol. 55, pp. 41-84.

Daily, L., Lovett, M., and Reder, L. (2001). Modeling individual differences in working memory performance: a source activation account. *Cognitive Science*, vol 25, pp 315–353.

Gilks, W.R., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Boca Raton, LA; Chapman and Hall/CRC.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.

Hodges, D. and Sargent, J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterized models. *Biometrika*, vol. 88, no. 2, pp. 367–379.

Holst, U. and Lindgren, G. (1991). Recursive estimation in mixture models with Markov regime. *IEEE Transactions on Information Theory*, vol. 37, no. 6, pp. 1683–1690.

Ito, H. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 324–333.

Johnson, R. and Wichern, D (1998) *Applied Multivariate Statistical Analysis*. Prentice Hall.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795.

Koedinger, K. and MacLarin, B. (2002) Developing a pedagogical domain theory of early algebra problem solving. Carnegie Mellon University Technical Report CMU-CS-01-119 (Computer Science) and CMU-HCII-02-100 (Human-Computer Interactions).

Langeheine, R. and Rost, J. (1988). *Latent Trait and Latent Class Models*. Plenum Press.

Lee, and Nelder, (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, series B*, vol. 58, no. 4, pp. 619–678.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, vol. 5, pp 81–91.

Moody, J. (1992). The effective number of parameters: analysis of generalization and regularization in nonlinear learning systems. pp 847-854 in Moody, Hanson and Lippman, *Advances in Neural Information Processing Systems 4*, Morgan Kauffmann Publishers, INC, 1992.

Myung, I. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, vol. 4, pp. 190–204.

Newell, A. (1973a). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. pp. 283–310 in W.G. Chase (ed.), *Visual information processing*. New York: Academic Press.

Newell, A. (1973b) Production systems: Models of control structures. pp. 463–526 in W.G. Chase (ed.), *Visual information processing*. New York: Academic Press.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge MS; Harvard University Press.

Petrov, A. (2001). Fitting the ANCHOR model to individual data: a case study in Bayesian methodology. pp. 174–180 in Altman, M., Cleeremans, A., Schunn, S., and Gray, W. (2001) *Proceedings of the 2001 International Conference on Cognitive Modeling*.

Pitt, M., Myung, I. and Zhang, S. (2002) Toward a method of selecting among competing models of cognition. *Psychological Review*, vol. 109, no. 3, pp. 472–491.

Raftery, A (1996). Hypothesis testing and model selection. Chapter 10 in Gilks, W.R., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Boca Raton, LA; Chapman and Hall/CRC.

Roberts, S. and Pashler, H (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, vol. 107, no. 2, pp. 358–367.

Salthouse, T. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological Bulletin*, vol 99, pp. 303–319.

Salvucci, D. and Anderson, J. (1998) Analogy. Chapter 10 in Anderson, J. and Lebiere, C. (1998) *The Atomic Components of Thought*. Lawrence Erlbaum Associates.

Schunn, C. and Wallach, D. (2001). Evaluating goodness of fit in comparison of models to data. Online manuscript available at

`http://www.lrdc.pitt.edu/schunn/gof/index.html`

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, vol 6, no. 2, pp. 461–464.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society series B*, Part 4, pp. 583–639.

Taatgen, N., and Anderson, J. (2002). Why do children learn to say "Broke"? A model of learning the past tense without feedback. *Cognition*, vol. 86, pp. 123–155.

Tanner, M. (1996). *Tools for Statistical Inference* (3rd edition). New York, Springer-Verlag.

Tanner, M. and Wong, W. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–540.

Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 120–131.